## REMARKS

Claims 10 and 12–17 are pending in this application. The Examiner has withdrawn from consideration non-elected claims 16 and 17. By this Amendment, claim 10 is amended. Support for the amendments to the claims may be found, for example, in the original claims and specification. No new matter is added.

In view of the foregoing amendments and following remarks, reconsideration and allowance are respectfully requested.

### I. Enablement Rejection Under 35 U.S.C. §112, First Paragraph

The Office Action rejects claims 10 and 12–15 under the enablement requirement of 35 U.S.C. §112, first paragraph. Applicants respectfully traverse the rejection.

### A. Claim Amendments

### 1. Scope Of "Patient"

The Office Action, at page 5, asserts the claims are drawn to any patient such as humans, dogs, and monkeys. The Office Action asserts that it would be unpredictable to associate the expression levels observed in one species to any other species, because such correlations cannot be directly extrapolated between species. Without conceding the propriety of the Office Action's reasoning, claim 10 is amended to recite "human patient."

### 2. Which Other 28 Genes Are Correlative To Prognosis?

The Office Action, at page 6, asserts that a skilled artisan would have to perform undue experimentation to determine which other 28 genes are correlative to prognosis. By this Amendment, claim 10 is amended to recite, "contacting the biological material with reagents specific for a combination of 9 to 37 target genes comprising the nucleic acid sequences set forth in SEQ ID NOs: 1–37, wherein the reagents include at least reagents specific for the target genes comprising the nucleic acid sequences set forth in SEQ ID NOs: 2, 3, 7, 8, 10, 22, 25, 29, and 34, respectively."

## B. The Office Action's Other Contentions

The Office Action presents several other contentions as to why the claimed methods are allegedly non-enabled. Among other things, the Office Action asserts:

- No stage 3 tumors were evaluated: "The instant specification has not provided an analysis in which the skilled artisan is provided guidance as to which group a patient with stage 3 should be placed."
- The claims encompass any expression level; and
- "It is unpredictable that the gene expression associations observed in the instant specification are reproducible."

The Office Action concludes that one of skill in the art would need to perform undue experimentation to determine any prognosis of neuroblastoma by detection of any level of expression of any of SEQ ID NOs:2, 3, 7, 8, 10, 22, 25, 29, or 34. Applicants respectfully disagree.

### 1. Only Two Prognoses Are Possible: "Good Prognosis" And "Poor Prognosis"

The Office Action, at page 12, asserts that one of skill in the art would have to perform undue experimentation to determine any prognosis of neuroblastoma. However, the claims allow for only two possibilities: (1) good prognosis, and (2) poor prognosis—there are not multiple different types of good prognoses nor multiple different types of poor prognoses. The method requires that once an expression profile for the patient is obtained, cluster analysis is performed on that expression profile against the expression profiles previously obtained from patients who were clinically classified as either good prognosis or poor prognosis. The claims define the criteria for the clinical classifications. Because of the nature of cluster analysis, the patient's expression profile will be clustered with either the previously classified poor prognosis patients or the previously classified good prognosis patients. It is an either/or situation—there are no fence sitters; there is no middle ground; there are no other categories or groups.

As discussed by Shannon et al., "Analyzing microarray data using cluster analysis," Pharmacogenomics (2003) 4(1), 41–52 ("Shannon," copy enclosed), the algorithms used in cluster analysis "are guaranteed to produce clusters from any data." Page 48, 1st column, 2nd full paragraph. Shannon explains in the preceding paragraph that in the absence of formal statistical tests, external criteria are typically used to choose the number of clusters, and that deciding which external criteria to use to define the clusters is subjective in nature. Although Shannon indicates that clustering analysis has some statistical pitfalls due to the subjective nature of cluster splitting and the fact that any data can be clustered, this does not detract from the fact that Shannon demonstrates that the claimed methods are indeed enabled.

For example, claim 10 defines the external criteria used to define the two groups or clusters. That is, the "good prognosis" cluster comprises the expression profiles from patients diagnosed with a stage 1, 2, or 4s neuroblastoma who did not die within 75 months of diagnosis, and the "poor prognosis" cluster comprises the expression profiles from patients diagnosed with a stage 4 neuroblastoma or who died within 75 months of diagnosis. The fact that no expression profiles from stage 3 neuroblastoma patients were used to define the clusters is completely irrelevant to the reproducibility of the claimed methods. If a sample from a patient clinically classified as having stage 3 neuroblastoma is subjected to the claimed method, the expression profile from that sample will be clustered with either the good prognosis group or the poor prognosis group.

Likewise, the fact that the claimed methods encompass any expression level does not require undue experimentation by one of skill in the art to practice the claimed invention. Again, as discussed by Shannon, any data can be clustered; therefore, any level of expression data can be clustered. Each sample is going to exhibit different expression levels for a given gene and, so, it is only logical that the claims encompass any expression level.

Furthermore, as shown in Table 2, the expression level for each gene is either underexpressed or overexpressed on average in poor prognosis patients compared to good prognosis patients. Table 2 identifies 13 genes with decreased expression and 24 genes with increased expression in poor prognosis patients. However, the expression profile of a given patient is not going to necessarily fit this 13/24 score. For example, a sample may exhibit decreased expression in 12/13 of the designated decreased expression genes for poor prognosis, while 1/13 genes exhibits increased expression. Clustering analysis will still be able to determine whether the expression profile should be clustered with the good prognosis patients or the poor prognosis patients.

### 2. Prognosis Does Not Necessarily Depend On The Age At Diagnosis Nor The Tumor Stage

The Office Action asserts that Ohira teaches that prognosis depends on the age at diagnosis and the tumor stage. This assertion is not entirely accurate. Ohira I (Cancer Letters (2005)) describes in its Introduction that poor prognosis of NBL patients depends on age at diagnosis (older than 1 year), advanced tumor stage (3 or 4), presence of *MYCN* amplification, low *TRKA* expression, unfavorable histology, diplody, and chromosomal loss of 1p36 in tumors. Certainly, Ohira I is not teaching that all seven of these factors must be present before a patient is determined to have a poor prognosis. When looking at this in context of the entire article, it is clear that Ohira I is merely providing background information about "conventional prognostic markers" (see page 6, 1st column, last paragraph). Indeed, Ohira I states that gene expression-based systems can predict prognosis "independently of currently known risk factors." Page 9, 2nd column, 1st full paragraph.

This is also clearly set forth in Ohira II (Cancer Cell (2005)). In the first full paragraph found in the first column on page 345, Ohira II teaches that when compared to other prognostic factors such as age at diagnosis, disease stage, and *MYCN* amplification, microarray analysis showed the best sensitivity-specificity balance for predicting the outcome of neuroblastoma independent of those other factors. Ohira II also teaches that if microarray analysis is combined with one or more other prognostic factors, accuracy can be further improved. *See* page 345, 1st column, 1st full paragraph. Nevertheless, contrary to the Office Action's assertions, Ohira discloses that microarray analysis may be used to accurately predict neuroblastoma outcome independent of the use of other traditional prognostic factors such as age at diagnosis and tumor stage. Furthermore, the claims do not exclude the use of other prognostic factors.

### 3. Unpredictability

The Office Action, at the paragraph bridging pages 9 and 10, states (citations omitted):

> The art teaches associations between expression studies and cancer prognosis are unpredictable and must be reproduced to determine if there is a correlation. Ohira et al. teaches a method of predicting prognosis of neuroblastoma using cDNA microarray. Ohira et al. teaches that gene expression analysis for cancer prognosis prediction should pay close attention to the reproducibility of obtained results. Ohira et al. teaches a complete cross validation analysis without introducing any information leakage and an independent test using new samples are necessary. Therefore Ohira et al. exemplify that validation of initial screening results is essential. Here in the instant case it is not clear if any of this analysis was undertaken therefore it is unpredictable whether the results observed are adequate basis for a prognostic too."

#### a) Ohira

Ohira II teaches that the reliability of prediction for outcomes of cancer patients is directly tied to the method's reproducibility. *See* page 338, 1st column, 1st full paragraph. Ohira II teaches that it is important to use "sound and highly reliable statistical methodologies." *Id.* Ohira II proposes that "a complete crossvalidation analysis without introducing any information leakage and an independent test are necessary." *Id.* Ohira II later clues the reader in to what is meant by "complete crossvalidation analysis" when it asserts that the microarray study reported by van't Veer et al. (2002) failed to include in their crossvalidation the validation of the number of genes used. *See* page 345, 1st column, last paragraph. Ohira I calculated that when a "complete validation was applied," van 't Veer's accuracy was actually 73.1% instead of the reported 80.7%.

The Office Action appears to imply that, in view of Ohira II, any microarray study that does not use a complete cross-validation analysis without introducing any information leakage and an independent test are unreliable/unpredictable. Applicants disagree.

Applicants would not consider a method that accurately predicts cancer outcome 73.1% of the time to be unreliable or unpredictable. Although Ohira II teaches that the accuracy of microarray analysis in cancer outcome prognosis may be improved by including in the statistical analysis validation of the number of genes used, it does not teach that microarray analysis without "complete crossvalidation" is unreliable or unpredictable.

Furthermore, as its "independent test," Ohira II discloses that they created a mini-chip system using the 200 top-ranked genes of their 5340 gene system, and ran 50 independent samples on the mini-chip system. *See* page 341, 2nd column. In other words, Ohira II used the same expression profiling platform for its mini-chip system as it did for its 5340 gene system. Interestingly, Schramm criticizes Ohira II (Reference 3 in Schramm) for not using a different expression profiling platform to independently validate its results. See Schramm, page 1463, 1st column, 2nd paragraph.

### b) Validation Used in Applicants' Disclosure

In response to the Office Action's assertion that "it is not clear if any of this analysis was undertaken" by the Applicants, Applicants respectfully submit that the specification discloses the following validation techniques at pages 26–32:

- To prevent the variations obtained by using various chips, an overall normalization approach was taken using the MAS5.0 software (Affymetrix), which converted the raw data obtained for each chip into an average signal within an intensity of 500. The results obtained on one chip could then be compared with the results obtained on another chip.

- The MAS5.0 software included a statistical algorithm to consider whether or not a gene was expressed. Each gene represented on the U95Av2 chip was covered with 16 to 20 pairs of probes of 25 oligonucleotides. The pairs of probes comprised a first probe that hybridized perfectly (PM, or perfect match, probe) with one of the cRNAs derived from a target gene, and a second probe, identical to the first probe with the exception of a mismatch (MM, or mismatched, probe) at the center of the probe. Each MM probe was used to estimate the background noise corresponding to a hybridization between two nucleotide fragments of non-complementary sequence. The specification cites to Affymetrix technical note "Statistical Algorithms Reference Guide"; and Lipshutz, et al. (1999) Nat. Genet. 1 Suppl., 20-24.

- The expression data analysis was carried out using Microsoft Excel, Spotfire Decision Site for Functional Genomics V7.1 software (Spotfire AB, Gothenburg, Sweden), and PAM (Prediction Analysis in Microarrays) module of the R statistics software (Ihaka & Gentleman (1996) Journal of Computational and Graphical Statistics 5, 299-314; Tibshirani, et al. (2002) Proc. Natl. Acad. Sci. 99, 6567-6572).

- Genes exhibiting an expression level that was comparable between all the groups of patients [Tibshirani, et al. Proc. Natl. Acad. Sci. 99, 6567-6572] were excluded. The genes that were nonexpressed in all the patients were also excluded (MAS5.0 software). Finally, some genes were excluded if the expression mean of the 2 groups (patients with a good prognosis and patients with a poor prognosis) was less than 500 or if the ratio of the expression means between the patients with a poor prognosis and the patients with a good prognosis was between 0.7 and 1.3.

- The expression of the remaining 1488 genes was subsequently analysed using the PAM algorithm (Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. 99, 6567-6572)[1] to arrive at a final count of 37 genes.

- The microarray results were validated using RT-PCR. The experiments were carried out in duplicate. *See* page 31. The quantification was carried out using the standard curve method and using the CT comparative method as recommended by the manufacturer. The standard curves were obtained using dilutions of cDNA from neuroblastoma cell lines, and were prepared for each PCR. The expression of the target gene was determined using these standard curves. The relative expression of each target gene was defined by comparison with the expression of the reference gene. The Pearson and Spearman correlation tests were used to calculate the correlation between the results obtained by microarray analysis and the results obtained by RT-PCR.

- 6 independent tumor samples from "test" patients without prior knowledge of their prognoses were tested and analyzed with respect to their expression profiles for the 37 genes, and were classified as good prognosis or poor prognosis after comparing the "test" expression profiles to the expression profiles previously obtained from the 23 clinically classified neuroblastoma patients. All 6 samples were later verified as being correctly classified. The 6 samples were also tested using smaller panels of 19 genes, 16 genes, 12 genes, and 9 genes. In each case, each test sample was correctly classified.

---

[1] Submitted with IDS filed May 31, 2006 (Reference 29); this article discusses in detail the statistical methodology and algorithms used by PAM, including the k-fold cross-validation it uses.

As outlined above, the specification is quite clear that a number of steps were taken to ensure the reproducibility of the claimed methods. Included in these steps are (1) cross-validation techniques, (2) validation by RT-PCR, which Schramm describes as the gold standard for measuring gene expression, and is considered to be the final proof of array data,[2] and (3) validation by running 6 independent test samples on 5 different gene panels.

### c) **Schramm**

The Office Action asserts that it is unpredictable that the gene expression associations observed in the Applicants' specification are reproducible because Schramm et al. discloses a group of genes that is predictive of prognosis but does not overlap the genes asserted by the instant specification. The Office Action asserts that despite the fact that Applicants' microarray data was validated by RT-PCR, which Schramm teaches is considered to be the **final proof of array data**, the Office Action, on page 17, indicates that Applicants have not produced sufficient "evidence that the expression data can be replicated in any individual patient."

It is unclear to Applicants what is meant by the phrase "evidence that the expression data can be replicated in any individual patient," but it appears that the Office Action is questioning either (1) whether expression data from any patient would result in the identification of the same 37 genes identified by the Applicants' disclosure, or (2) whether the claimed methods can be reliably practiced for any patient. In either case, it appears that the Office Action bases its questioning on the fact that "Schramm et al. uses the same microarray as used in the instant specification and did not find associations between the 9 genes claimed in the instant specification and prognosis." *See* Office Action, page 17.

---

[2] *See also* Takita, page 125, 2nd column, which indicates that good correspondence between RT-PCR analysis and array-based determinations demonstrates that the array-based determination are highly reproducible.

In response, Applicants again respectfully point out that Schramm indicates that "several genome-wide mRNA expression profiling studies have identified reliable outcome predictors for neuroblastoma, but with little or no overlap in the decision-making genes" (emphasis added). *See* page 1, 1st column, 3rd sentence. In other words, Schramm states that despite the fact that there was little or no overlap between the three studies, the results of each study were still considered reliable. Thus, based on this, just because there is no overlap between the Applicants' claimed decision-making genes and those disclosed by Schramm does not mean that either disclosure is unreliable.

Also, Schramm's classifier was different from the Applicants' classifier. Schramm discloses that it obtained, by using Affymetrix U95A chips and PAM, a set of 39 genes able to discriminate between neuroblastoma patients with recurrent tumors and those with no evidence of disease following initial therapy. *See* page 1460, 2nd column, 2nd paragraph. Because Schramm's classifier was entirely different from Applicants' classifier, it is not unexpected that Schramm's decision-making genes are different from Applicants' decision-making genes.

Ein-Dor et al., "Outcome signature genes in breast cancer: is there a unique set?", Bioinformatics 21(2):171–178 (2005) ("Ein-Dor," copy enclosed), identifies a similar phenomenon in three studies that identified sets of breast cancer survival-related genes that had only a few genes in common. *See* Ein-Dor, Introduction. All three microarray studies yielded gene sets whose expression profiles successfully predicted survival in breast cancer.

*Id.* at page 173, 1st column, 3rd paragraph. Ein-Dor indicates that this lack of agreement can be attributed to different chips; different methods of: sample preparations, mRNA extraction, and analysis of the data; and "most importantly" to genuine differences between the patients such as tumor grade, stage, etc. *Id.*

To eliminate the sources of variation, Ein-Dor focused on data from a single experiment (van't Veer et al., 2002) that involved 96 patient samples (77 training samples and 19 test samples) and 5852 genes. After ranking all the genes according to their correlation with survival, the study built a series of classifiers based on consecutive groups of 70 genes, and found seven other sets of 70 genes with the same prognostic capabilities of those based on the top 70 genes (using the methods and training set used by van't Veer). To ensure that their results were not unique to the specific training and test sets selected by van't Veer, Ein-Dor repeated the procedures for 1000 different compositions of the 77 training sets and the 19 test sets (i.e. different patient pools). Each training set was used to rank the genes, and for each case the sequence of classifiers previously described was constructed and analyzed. The results showed that for each of the training sets, classifiers based on very low-ranked genes were capable of predicting survival with quality similar to the high-ranking genes. *See* page 173, "Results" section.

Ein-Dor concluded that the relative ranking of genes on the basis of correlation with survival changes drastically when a different training set (i.e. pool of patient samples) is used. *See p*age 177, "Discussion" section. Ein-Dor states,

> "These large fluctuations in gene rank indicate that the identities of the top 70 range of genes are not robust, and hence will not be reproduced in a different experiment.... A high sensitivity of the results to the arbitrary decisions [e.g. choice of training and tests set] may indicate that the conclusions, e.g. the list of survival-related genes, are not unequivocal. In light of the inconsistency between lists of survival-related genes generated from the same dataset, the disagreement between lists obtained from different datasets is not surprising. A

> possible biological explanation for this may be the individual variations and heterogeneities associated with markers for outcome, even within a clinically homogeneous group of patients."

*Id.*

Therefore, if the Office Action is questioning whether expression data from any patient (or more correctly, any set of patients) would result in the identification of the same 37 genes identified by the Applicants' disclosure, Ein-Dor would indicate no. However, if the Office Action is questioning whether Applicants' claimed methods may be reliably used to determine a good or poor prognosis for as claimed for any neuroblastoma patient, Ein-Dor would indicate yes.

In particular, Ein-Dor indicates that one should separate two issues: (1) the quest for survival-related master genes, and (2) the construction of prognostic tools on the basis of a short gene list. *Id.* Ein-Dor teaches, "One can produce fairly reliable prognostic tools; many genes are related to survival, and using a large enough subset of them will compensate for the fluctuations in the predictive power of individual genes for individual patients. Membership in a prognostic list, however, is not necessarily indicative of the gene's importance in cancer pathology." *Id.*

### C. Conclusion

In summary, the above discussion demonstrates that the claimed methods are fully enabled. The disclosure, coupled with the knowledge that was readily available at the time of invention, enables one of skill in the art to obtain an expression profile from a patient, and to perform cluster analysis of the patient's expression profile with expression profiles of the same target genes from patients previously clinically classified as good prognosis and expression profiles of the target genes from patients previously clinically classified as poor prognosis (according to the definitions of good prognosis and poor prognosis set forth in the claims). Because of the nature of clustering analysis, the method will result in either "good prognosis" or "poor prognosis" regardless of the patient's age, tumor stage, or expression levels.

The question thus becomes one of operability—the Office Action questions whether the claimed methods can reliably or accurately predict "good prognosis" or "poor prognosis" with any patient. But operability is a question of utility. MPEP §2164.07 states that where the subject matter of a claim has been shown to be non-useful or inoperative, the Examiner should make a rejection under 35 U.S.C. §112, first paragraph, and a rejection under 35 U.S.C. §101. However, MPEP §2164.07 warns that Office personnel should not impose a 35 U.S.C. §112, first paragraph, rejection grounded on a "lack of utility" basis (e.g. inoperability) unless a 35 U.S.C. §101 rejection is proper. "In particular, the factual showing needed to impose a rejection under 35 U.S.C. 101 must be provided if a 35 U.S.C. 112, first paragraph, rejection is to be imposed on 'lack of utility' grounds."

The Office Action has not made a rejection under 35 U.S.C. §101, nor has it provided a reasonable basis to support a conclusion that the claimed methods are inoperable. The Office Action has not indicated what portions of the Applicants' disclosed methodologies would cause one of skill in the art to doubt the operability of the claimed methods. The fact that different studies have produced sets of decision-making genes with little or no overlap does not establish that a given set of decision-making genes is inoperable.

Accordingly, Applicants respectfully request reconsideration and withdrawal of the rejection.

## II. Rejoinder

Applicants respectfully request rejoinder of claims 16 and 17. Claims 16 and 17, although requiring a larger combination of target genes than the elected combination of target genes, require in their combinations the elected combination as a subset. Therefore, if claim

10 is found allowable, then claims 16 and 17 should also be found allowable, as they require all the limitations of claim 10.

## III. Conclusion

In view of the foregoing, it is respectfully submitted that this application is in condition for allowance. Favorable reconsideration and prompt allowance of the application are earnestly solicited.

Should the Examiner believe that anything further would be desirable in order to place this application in even better condition for allowance, the Examiner is invited to contact the undersigned at the telephone number set forth below.

Respectfully submitted,

William P. Berridge
Registration No. 30,024

Jeffrey R. Bousquet
Registration No. 57,771

WPB:JRB

Attachments:
Shannon et al., "Analyzing microarray data using cluster analysis", Pharmacogenomics (2003) 4(1), 41–52
Ein-Dor et al., "Outcome signature genes in breast cancer: is there a unique set?", Bioinformatics 21(2):171–178 (2005)

Date: November 23, 2009

**Oliff & Berridge, PLC**
**P.O. Box 320850**
**Alexandria, Virginia 22320-4850**
**Telephone: (703) 836-6400**

DEPOSIT ACCOUNT USE
AUTHORIZATION
Please grant any extension
necessary for entry;
Charge any fee due to our
Deposit Account No. 15-0461

# Analyzing microarray data using cluster analysis

*William Shannon*†1,2, *Robert Culverhouse*1 & *Jill Duncan*1

†*Author for correspondence*
1*Department of Medicine*
2*Division of Biostatistics, Washington Univ. School of Medicine, 660 S. Euclid Ave, Campus Box 8005, St. Louis, MO 63110, USA*
*Tel: +1 314 454 8356;*
*Fax: +1 314 454 5113;*
*E-mail: shannon@ilya.wustl.edu*

**As pharmacogenetics researchers gather more detailed and complex data on gene polymorphisms that effect drug metabolizing enzymes, drug target receptors and drug transporters, they will need access to advanced statistical tools to mine that data. These tools include approaches from classical biostatistics, such as logistic regression or linear discriminant analysis, and supervised learning methods from computer science, such as support vector machines and artificial neural networks. In this review, we present an overview of another class of models, cluster analysis, which will likely be less familiar to pharmacogenetics researchers. Cluster analysis is used to analyze data that is not *a priori* known to contain any specific subgroups. The goal is to use the data itself to identify meaningful or informative subgroups. Specifically, we will focus on demonstrating the use of distance-based methods of hierarchical clustering to analyze gene expression data.**

## Introduction

As gene chips become more routine in basic research, it is important for biologists to understand the biostatistical methods used to analyze these data so that they can better interpret the biological meaning of the results. Strategies for analyzing gene chip data can be broadly grouped into two categories: *discrimination* (or *supervised learning*) and *clustering* (or *unsupervised learning*).

Discrimination requires that the data consist of two components. The first is the gene expression measurements from the chips run on a set of samples. The second component is the data characterizing the samples (e.g., tumor or normal tissue, time cells were harvested from a culture) or the genes (e.g., regulatory factor, oncogene). For this method, the goal is to use a mathematical model to predict a sample characteristic, say tumor subtype, from the expression values. Once this model is fit, the gene expression values of a new tumor sample can be used to make a 'prediction' of its subtype class. There are a large number of statistical and computational approaches for discrimination (i.e., supervised learning) ranging from classical statistical linear discriminant analysis [1] to modern machine learning approaches such as support vector machines [2,3] and artificial neural networks [4,5]. Microarray analysis using supervised learning methods was recently reviewed in this journal [6] and will not be discussed further in this review.

In this review, we will discuss the second group of analytical approaches for analyzing microarray data: *cluster analysis* or *unsupervised learning*. In clustering, the data consist only of the gene expression values. The analytical goal is to find clusters of samples or clusters of genes such that observations within a cluster are more similar to each other than they are to observations in different clusters. Cluster analysis can be viewed as a data reduction method in that the observations in a cluster can be represented by an 'average' of the observations in that cluster.

There are a large number of statistical and computational approaches available for clustering. These include hierarchical clustering [7,8] and k-means clustering [9] from the statistical literature and self-organizing maps [10] and artificial neural networks [4] from the machine learning literature. While these algorithms are relatively equivalent in terms of performance (i.e., one method does not dominate all others), the focus of this paper will be on hierarchical clustering. For a broad overview of the multivariate statistics used in cluster analysis the reader is referred to Timm [11]. For a broad overview of both unsupervised and supervised learning methods from both the statistics and machine learning literature, the reader is referred to Hastie *et al.* [12]. For a broad overview of the application of these methods to biological data the reader is referred to Legendre and Legendre [13]. Each of these references cover hierarchical and other clustering methods in more mathematical detail than presented here and show their application to data for illustration.

**Keywords:** consensus methods, distance calculations, heat maps, hierarchical clustering, k-means clustering, Mantel statistics, microarrays, unsupervised learning

*Ashley Publications Ltd*
*www.ashley-pub.com*

### Raw data

Gene expression data measured by gene chips (microarrays) are preprocessed using image analysis techniques to extract expression values from images and scaling algorithms to make expression values comparable across chips. These preprocessing steps are generally done with a microarray-platform vendor's software or through software developed by researchers interested in improving the estimates of the expression data [14,15]. While these steps can have a significant impact on the quality of the data and are an area of active research, our review will start with the assumption that these preprocessing steps have already been performed and the estimates of the expression level are as good as can be obtained.

Expression data are typically analyzed in matrix form with each row representing a gene and each column representing a chip or sample. For a study with 20 samples run on Affymetrix GeneChips™, the dimensions of the data matrix would be (approximately) 12,000 rows (one for each gene) by 20 columns. Newer chips have even more genes on them. Often there will be one additional column giving the gene label for identification. However, this column is excluded from analysis and only the chip columns containing expression values are used.

We represent the data matrix by the symbol X and denote the data as follows:

$$X = \begin{array}{c|ccccc}
\text{Gene} & \text{Chip1} & \text{Chip2} & \cdots & \text{Chip20} \\
\hline
1 & x_{1,1} & x_{1,2} & \cdots & x_{1,20} \\
2 & x_{2,1} & x_{2,2} & \cdots & x_{2,20} \\
3 & x_{3,1} & x_{3,2} & \cdots & x_{3,20} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
12{,}000 & x_{12000,1} & x_{12000,2} & \cdots & x_{12000,20}
\end{array}$$

The matrix entries correspond to the expression value of a gene (row) and chip (column). For example, $x_{1,1}$ is the expression value of gene 1 in sample 1, $x_{3,20}$ is the expression value of gene 3 in sample 20, etc. In general the notation $x_{i,j}$ corresponds to the expression level of gene $i$ in sample $j$. While this notation may seem clumsy at first, it is important to understand the 'structure' of the data to learn how the analysis is done and how the results should be interpreted.

Most software programs use the data matrix $X$ in this form to cluster genes. There is no reason that clustering cannot also be done directly on columns. However, to simplify discussion in this paper and to be consistent with many statistical packages, to cluster samples we will use the *transposition* of $X$. This is obtained by flipping the matrix across the diagonal so that the columns become the rows and the rows become the columns. This changes the dimensions from the original 12,000 rows by 20 columns to a matrix of dimension 20 rows by 12,000 columns. In this format the samples are the rows and the genes are the columns. We denote the transposition of $X$ by $X^T$:

$$X^T = \begin{array}{c|cccccc}
\text{Chip} & \text{Gene 1} & \text{Gene 2} & \text{Gene 3} & \cdots & \text{Gene 12000} \\
\hline
1 & y_{1,1} & y_{1,2} & y_{1,3} & \cdots & y_{1,12000} \\
2 & y_{2,1} & y_{2,2} & y_{2,3} & \cdots & y_{2,12000} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
20 & y_{20,1} & y_{20,2} & y_{20,3} & \cdots & y_{20,12000}
\end{array}$$

The matrix entries of $X^T$, coded as $y_{i,j}$, correspond to the expression value in a chip (row) for a given gene (column). For example, $y_{3,20}$ is the expression value in sample 3 for gene 20. In general, the notation $y_{i,j}$ corresponds to the expression level in sample $i$ for gene $j$, so $y_{i,j} = x_{j,i}$.

### Filtering

The first step in analyzing microarray data is to filter out genes that are not expressed or do not show variation across sample types. In our experience, this usually reduces the data set by 3000–5000 genes. The Affymetrix GeneChips contains a variable for each gene that declares whether the gene was expressed, not expressed or indeterminate. We always remove from the analyses the rows corresponding to genes that were not expressed on any of the chips. Other strategies for gene filtering include filtering at a threshold of the variance of the gene across chips or if two or more tissue types are represented in the experiments, filtering at a threshold of a test statistic. For example, if gene chips are used to analyze tumor and normal tissues, the two groups can be compared using t-statistics calculated for each gene. An arbitrary threshold based on a value for the t-statistic or to filter out a certain percentage of the genes can be used.

These methods of filtering genes are arbitrary (except, perhaps, for filtering out genes based on the expression/no expression call by Affymetrix software). However, if used conservatively to filter out only the least differentially expressed genes, the analyst should be protected from eliminating any important genes.

### Standardized data

Although clustering methods can be applied to the raw data, it is often more useful to precede the analysis by standardizing the expression values. Standardization in statistics is a commonly used tool to transform data into a format needed for meaningful statistical analysis [16]. For example, ***variance stabilization*** is needed to fit a regression model to data where the variance for some values of the outcome $Y$ may be large, say for those values of $Y$ corresponding to large values of the predictor variable $X$, while the variance of $Y$ is small for those values corresponding to small values of $X$. Another use of standardization is to ***normalize*** the data so a simple statistical test (e.g., t-test) can be used. Transformations specifically designed to allow standard statistical tests to be applied to microarray data are currently being proposed [17,18].

Transformation of microarray data for cluster analysis has a different purpose than transformations used to meet assumptions of statistical tests as described above. Cluster analysis depends on a distance measure (discussed in the next section). Since distance measures are sensitive to differences in the absolute values of the expression values (scale), microarray data for clustering often needs to be transformed to adjust for different scales. To illustrate this, consider three hypothetical genes A, B and C, whose expression levels have been measured in four normal tissue samples and four diseased tissue samples. The results of these measurements are displayed in Figure 1A. Genes A and B are tightly coregulated and differentially expressed across tissue types (i.e., higher in diseased tissue relative to normal tissue) but gene A is expressed at a much higher level than gene B. Gene C is not differentially expressed across tissue types but happens to have average expression levels similar to that of gene A. We typically want to find clusters that place genes A and B together because they appear to be coregulated (low in normal tissue, high in diseased tissue) but would not cluster them with gene C which is constant across all tissue samples. Clustering using the raw expression profiles would separate genes A and B and cluster genes A and C. Figure 1B shows the expression profiles for the same three genes after normalization (see below) across samples. In this transformed data, we see the expression values for genes A and B are closely aligned. In contrast, the values for gene C fluctuate randomly. This transformation results in the representations for genes A and B being near each other and thus increases the likelihood that they are clustered together.

Normalizing a gene across samples is accomplished by subtracting from each expression level the mean of the expression levels for that gene and then dividing by the standard deviation of that gene. Our matrix notation in the last section can now be used to clarify how the normalization is done. The data matrix $X$ consists of rows of genes we want to normalize. Consider the first gene at row 1 consisting of the expression levels $x_{1,1}, x_{1,2}, \ldots, x_{1,20}$ corresponding to gene 1 in sample 1, gene 1 in sample 2 etc. We calculate the mean of gene 1 by

$$\bar{x}_1 = \frac{x_{1,1} + x_{1,2} + \ldots + x_{1,20}}{20}$$

and the standard deviation of gene 1 by

$$s_1 = \sqrt{\frac{(x_{1,1} - \bar{x}_1)^2 + (x_{1,2} - \bar{x}_1)^2 + \ldots + (x_{1,20} - \bar{x}_1)^2}{20 - 1}}$$

The notation '+ … +' indicates to add all the terms between $x_{1,2}$ and $x_{1,20}$. With these terms, the normalized expression values are:

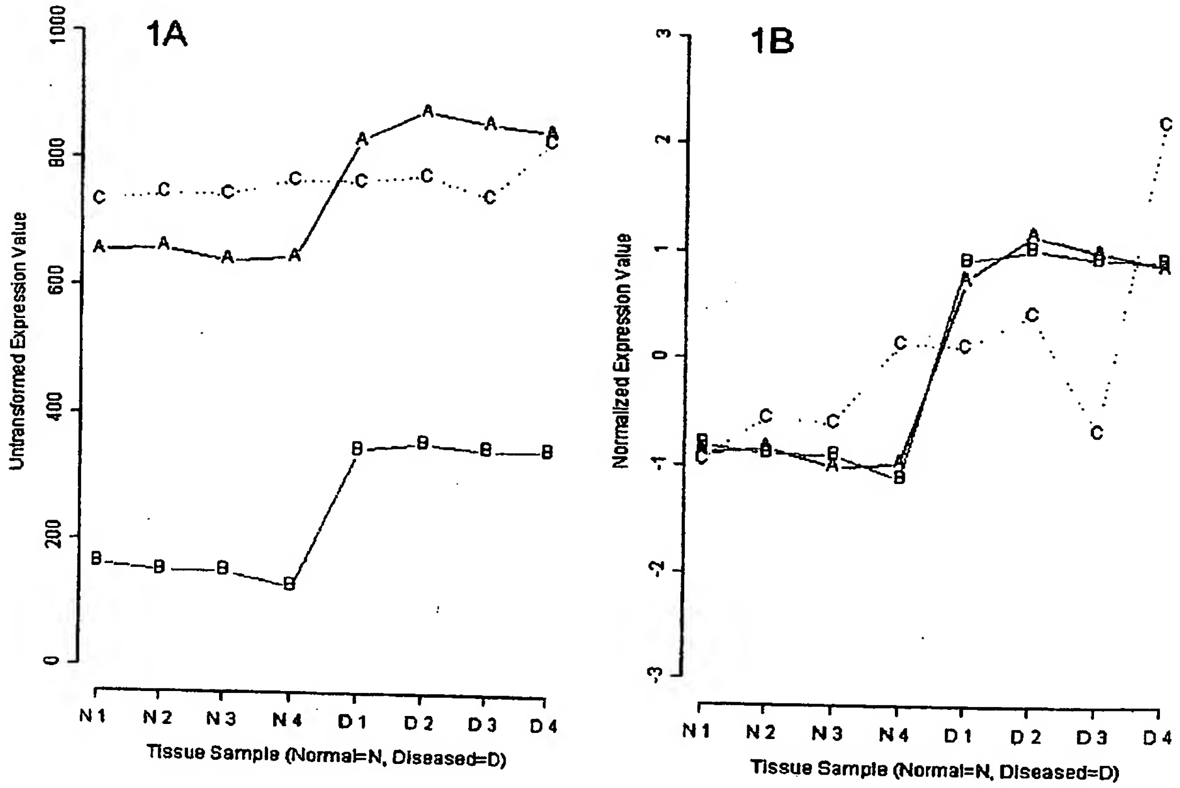
$$\frac{x_{1,1} - \bar{x}_1}{s_1}, \frac{x_{1,2} - \bar{x}_1}{s_1}, \ldots, \frac{x_{1,20} - \bar{x}_1}{s_1}$$

Most statistical programs can apply this normalization to each row of the matrix $X$. While conceptually this normalization might also be applied to each row of the transposed matrix $X^T$, we have found this not useful for uncovering structure in cluster analysis. We recommend that normalization be applied to genes across samples only.

### Distance measures

Many cluster analysis methods, including hierarchical clustering, use distances measured between rows of the data matrices $X$ or $X^T$. Measuring distances can be thought of as placing a ruler between two points and recording how far apart they are. To make this idea more clear before we present formulae for calculating distances, consider a simple example where the data matrix $X$ consists of gene expression values (rows) measured on only two samples (columns).

Figure 1. Gene expression profiles before and after normalization.



With just two samples (chips) we can plot each gene as a point on a two-dimensional scatter plot where the X-axis corresponds to the first chip and the Y-axis corresponds to the second chip. Consider three genes (A, B and C) whose expression levels are measured in the two samples:

| Gene | Chip1 | Chip2 |
|---|---|---|
| A | -2.0 | 1.0 |
| B | -1.5 | -0.5 |
| C | 1.0 | 0.25 |

These three genes can be plotted on a standard scatter plot as shown in Figure 2.

In addition to the gene labels A, B and C, this graph shows the calculated distances between each of these genes where the distances are calculated using the Euclidean distance formula. Specifically the distance between genes A and B is calculated by the formula

$$d(A,B) = \sqrt{(-2.0-(-1.5))^2+(1.0-(-0.5))^2} = 1.58$$

between genes A and C by

$$d(A,C) = \sqrt{(-2.0-1.0)^2+(1.0-0.25)^2} = 3.09$$

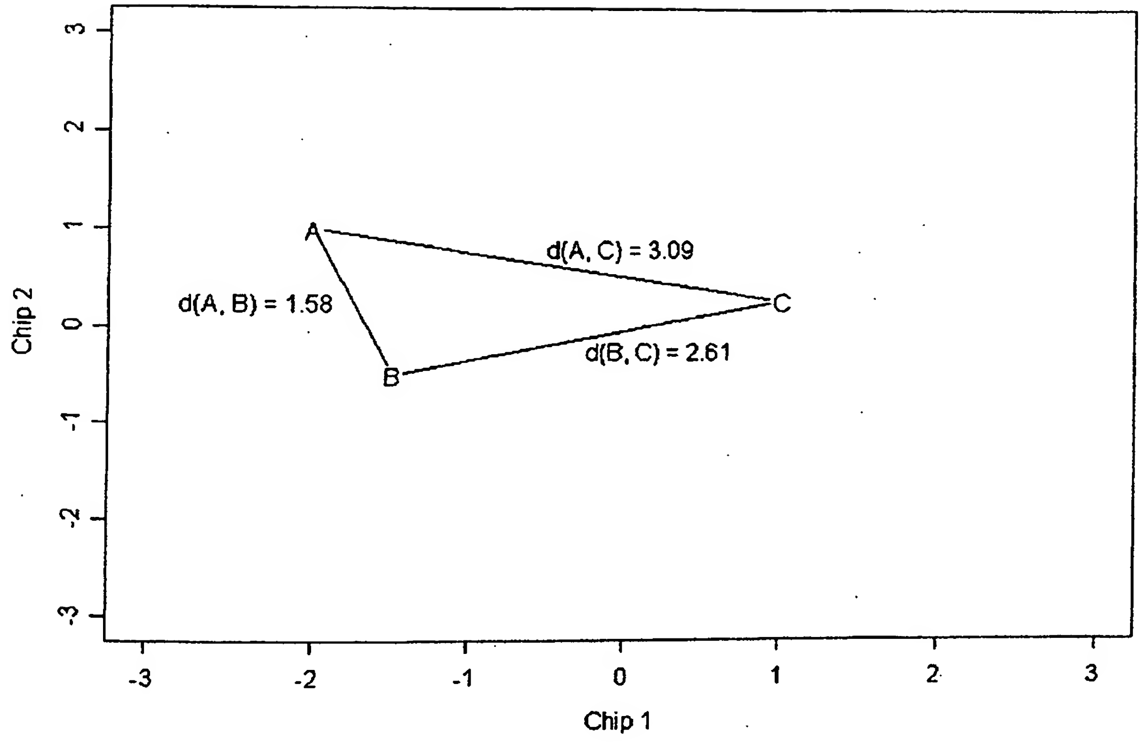
and between genes B and C by

$$d(B,C) = \sqrt{(-1.5-1.0)^2+(-0.5-0.25)^2} = 2.61$$

For convenience we record distances in a distance matrix

$$D = \begin{array}{cc} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \begin{bmatrix} 0.00 & 1.58 & 3.09 \\ 1.58 & 0.00 & 2.61 \\ 3.09 & 2.61 & 0.00 \end{bmatrix} \end{array}$$

Figure 2. Scatterplot of three genes from two samples.



The entries correspond to the distances between the genes denoted on the row and column (e.g., d(A,B) = 1.58). Note that the distances on the diagonal are all 0, the distances are all non-negative and the matrix is symmetric (e.g., d(A,B) = d(B,A)).

We now want to generalize the idea of the Euclidean distance matrix for any microarray data. Specifically, recall the data matrix is

$$X = \begin{array}{c|ccccc} \text{Gene} & \text{Chip1} & \text{Chip2} & \dots & \text{Chip20} \\ \hline 1 & x_{1,1} & x_{1,2} & \dots & x_{1,20} \\ 2 & x_{2,1} & x_{2,2} & \dots & x_{2,20} \\ 3 & x_{3,1} & x_{3,2} & \dots & x_{3,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 12{,}000 & x_{12000,1} & x_{12000,2} & \dots & x_{12000,20} \end{array}$$

where each row represents a gene and each column represents a chip. We will assume the data has been normalized. Distances are calculated between each pair of rows in X: d(1,2) is the distance from row 1 to 2, d(1,3) is the distance from row 1 to 3, d(1,12000) the distance from 1 to 12000, d(2,3) the distance from row 2 to 3, etc. The Euclidean distance for any two rows, say rows $i$ and $j$, is calculated using the expression values for all the chips in those two rows as follows:

$$d(i,j) = \sqrt{(x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2 + \dots + (x_{i,20} - x_{j,20})^2}$$

Notice that the subscripts on the $x$'s change for the column (chip) number. In words, this calculation is performed by subtracting the expression level of gene $j$ from gene $i$ and squaring it in each chip from 1 to 20, adding these values together and then taking the square root of the sum. No matter how many chips there are, for 12,000 genes this produces a 12,000 by 12,000 distance matrix containing 144,000,000 numbers indicating the computational complexity involved in cluster analysis.

There are many other distance measures that could be used (i.e., Manhattan distance) though we believe the Euclidean distance is generally appropriate for normalized microarray data.

### Hierarchical clustering

Several different algorithms will produce a hierarchical clustering from a pair-wise distance matrix. Each of these algorithms follows the same general strategy. Suppose we are clustering genes. The algorithms begin with each gene by itself in a separate cluster. These clusters correspond to the tips of the *clustering tree* (dendrogram). The algorithms search the distance matrix for the pair of genes that have the smallest distance between them and merge these two genes into a cluster. The distance matrix is recalculated to now include the distance between genes not clustered and the new cluster formed by the two genes. For simplicity, we will assume that only two genes are merged at each step, though more could be merged at any step.

Many algorithms follow this series of steps to produce hierarchical clustering of data. Variations between the algorithms can lead to different dendrograms and hence different clusters. We will consider an *average linkage* algorithm called *unweighted centroid clustering* for illustration and then compare it to other hierarchical clustering algorithms. It should be noted that different authors define average clustering in different ways. For example, others refer to the definition of average clustering used by Hastie *et al.* [12] as *unweighted arithmetic average clustering*. Readers interested in more technical descriptions of four average clustering algorithms should refer to Legendre and Legendre [13].

To illustrate our average linkage algorithm, recall the distance matrix calculated above for three genes A, B and C. Suppose we have added a fourth gene D and recalculated the distance matrix $D$,

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array}\!\!\begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.00 & 1.58 & 3.09 & 4.74 \\ 1.58 & 0.00 & 2.61 & 5.00 \\ 3.09 & 2.61 & 0.00 & 2.70 \\ 4.74 & 5.00 & 2.70 & 0.00 \end{bmatrix} \end{array}$$

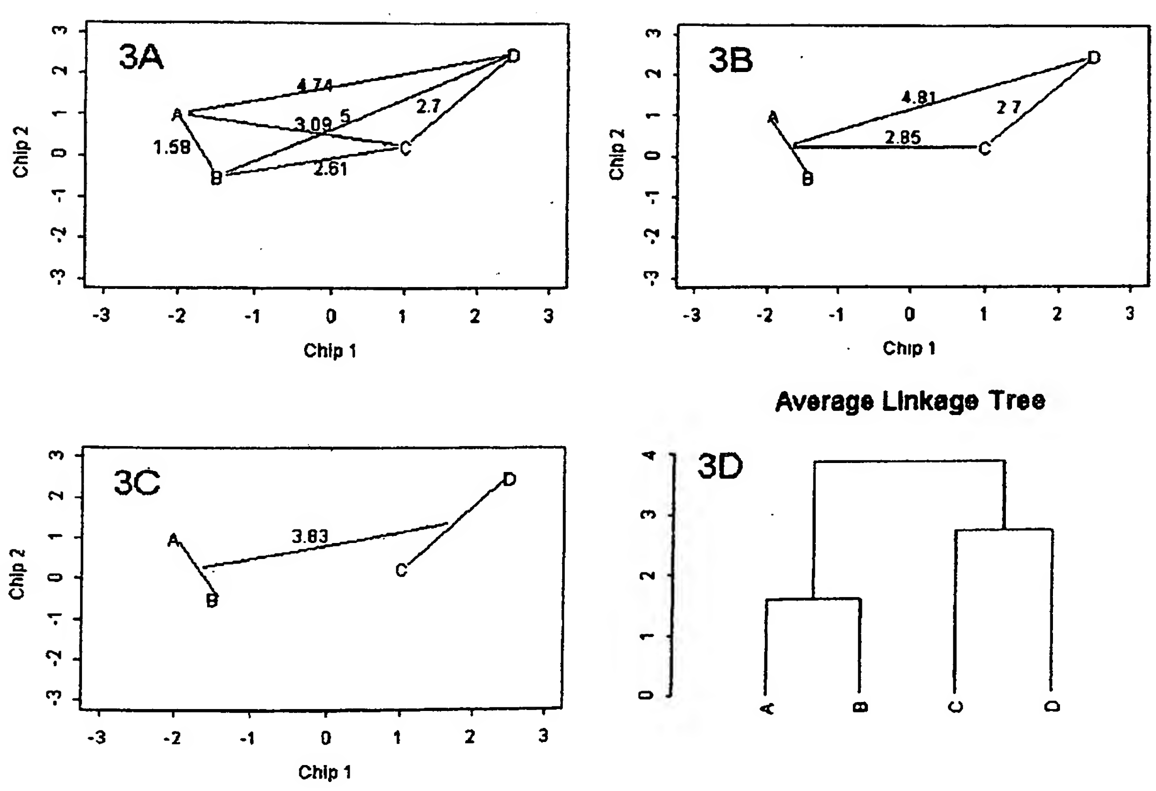
Figure 3A–D shows the steps of the average linkage clustering and the dendrogram obtained. In Figure 3A the four genes are plotted and the distance between each pair is indicated on the line connecting them. Initially, the algorithm finds the pair of genes closest to each other and merges them into a cluster. For this example, the first step merges genes A and B whose distance is 1.58. The distances are updated as follows: Replace the two genes A and B by the midpoint (AB) between them and recalculate the distance of gene C to this midpoint (d(AB, C) = 2.85) and gene D to this midpoint (d(AB, D) = 4.81). Note that d(C, D) = 2.7 is unchanged. The updated distances are shown in Figure 3B. The algorithm then repeats by finding the genes (or clusters) that have the smallest distance between them. In this iteration, genes C and D are clustered and replaced by their midpoint. The distance to all other gene clusters (such as AB) from this midpoint is calculated and the algorithm is repeated. Figure 3C shows the final distance d(AB, CD) = 3.83. Gene clusters AB and CD are merged in the last step of the algorithm.

Figure 3D summarizes the results of applying the average linkage algorithm to this data in a single graph known as a *dendrogram*. Initially, the four genes A, B, C and D are represented as single clusters along the bottom of the plot. Genes A and B are merged first at the level 1.58, followed by genes C and D being merged at the level 2.7, followed by AB and CD being merged at level 3.83. The dendrogram was fit and displayed using S-Plus (Seattle, Washington) software.

We presented the average linkage algorithm in Figure 3 in two ways to emphasize the relationship between a dendrogram and the pair-wise spatial distances of the genes. In this example, there were two chips so that the genes could be plotted in a scatter plot but the concept is the same for experiments with more than two chips. In general, the genes are points in a space whose number of dimensions equals the number of chips in the experiment. Regardless of the dimension of the problem, the Euclidean distance between genes or gene clusters can be calculated and each iteration of the algorithm merges the genes or gene clusters that have the smallest distance.

The final clustering of the genes is determined by where the dendrogram is cut. For example, cutting the dendrogram at level 3 (on the y-axis) results in the two clusters AB and CD, while cutting the dendrogram at the level 2 produces the three clusters AB, C and D. This dendrogram

Figure 3. Iterations of a hierarchical clustering and the resulting dendrogram.



can produce four distinct cluster results: ABCD when the dendrogram is not split; AB and CD when split into two groups; AB, C and D when split into three groups; and A, B, C and D when split into four groups.

Average linkage is one of many hierarchical clustering algorithms that operate by iteratively merging the genes or gene clusters with the smallest distance between them followed by an updating of the distance matrix. Many of these differ only in how the distance matrix is updated. In average linkage, as shown above, when two genes are clustered, the distances of the other genes and gene clusters to this new cluster is based on the midpoint of the new cluster. In contrast, single linkage calculates the distances between each gene in the new cluster to each of the genes in another cluster and takes the smallest distance. Complete linkage uses the largest distance of all these distances as the distance between the clusters. For example, in Figure 3A the first merging clustered genes A and B and the distance of this new cluster to gene D was d(AB, D) = 4.81. For single linkage, the distance would be d(AB, D) = 4.74 and for complete linkage the distance would be d(AB, D) = 5.

In practice, we have found the average linkage algorithm generally works well with standardized microarray data and single linkage generally performs poorly.

## Difficulties and pitfalls of cluster analysis

Unlike standard statistical methods, such as the t-test and analysis-of-variance, hierarchical clustering does not have a probabilistic foundation. Because of this, hierarchical clustering has no statistical test to guide the decision of where to cut the dendrogram. While it is possible to compute a formal test statistic, such as an F-test statistic, the assumptions of the statistical test are not met. Thus, the p-value listed in a statistics table would not represent the probability of the test-statistic value arising under the null hypothesis. In other words, the p-value has no meaning and is not a measure of the statistical significance of the clusters being different.

In the absence of formal statistical tests, external criteria are typically used to choose the number of clusters. One such criterion is that if splitting a tree at a particular point produces clusters of genes or samples that are nearly homogeneous with regard to an important property, the split would be deemed appropriate. For example, if splitting a tree at a particular height resulted in mostly tumor samples in one cluster and mostly normal samples in the other, the split would likely be considered interesting. Such a split is considered to be evidence that some of the genes used to generate the tree may be involved with the biology of the tumor and hence the genes warrant further scrutiny. The obvious problem with this approach is the subjective nature of deciding which external criteria to use.

A second difficulty with cluster analysis is that the algorithms are guaranteed to produce clusters from any data and there is currently no generally accepted way to test a null hypothesis of no clusters (e.g., data are distributed uniformly). For this reason, caution is required in interpreting the results of a cluster analysis method. The results always need to be examined to see if it is plausible that they are indeed natural clusters and not just artifacts of the algorithm.

In spite of these two problems, cluster analysis is a powerful tool for data reduction. One must remember that data reduction is the chief purpose of a cluster analysis. Since microarrays present the researcher with thousands of gene expression values, the data must be reduced before a human can tell an explanatory story about the relationship between genes and the phenotypes. Putative relationships between clusters of genes and phenotypes need to be recognized as nothing more than hypotheses generated by clustering methods. The clustering process has not statistically validated the relationships and they must be formally validated through additional experiments.

Methods are currently being developed to address the weaknesses of cluster analysis. We believe that the current interest in applying cluster analysis to genomics will generate enough research effort to successfully meet this challenge. For example, one method to determine the number of clusters without resorting to external criteria is to use the number that optimizes the Gap statistic, a statistic comparing within-cluster dispersion (spread of data points) to dispersion under the null hypothesis [12,19]. Another approach uses a perturbation method (sensitivity analysis). It introduces a small amount of random noise to the expression data, reclusters the data and then compares the results to the original clustering [20]. Our lab has begun investigating a formal statistical approach based on graph theory and a probability distribution on graphical objects [21] and another approach based on Mantel statistics which are briefly discussed below [22]. In spite of these efforts, the problem of selecting the correct number of clusters remains open after fifty years of study.

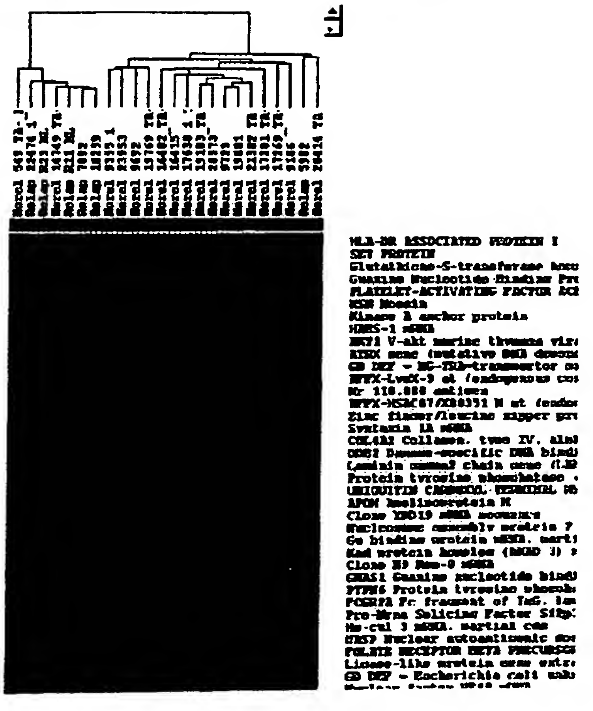
In summary, in spite of the danger of misusing or misinterpreting the results of cluster analysis, as long as one keeps in mind that cluster analysis is only appropriate for data reduction and hypothesis generation, the pitfalls can be reduced or avoided.

## Heat maps

Hierarchical clustering is used to produce what have been called 'heat maps' in papers reporting on microarray data analyses. The heat map presents a grid of colored points where each color represents a gene expression value in the sample. Figure 4 is an example taken from a recent paper using expression levels for cancer classification [23]. The grid coordinates correspond to the sample by gene combinations. In this case, the columns (samples) are tumors, some from patients who have relapsed and some from patients who have not relapsed. The rows represent 348 genes found to distinguish the patients according to their relapse status. In the heat map colors at a particular point (i.e., row by column coordinate) are assigned to represent the level of expression for that gene (row) in the sample (column) with red corresponding to high expression, green corresponding to low expression and black corresponding to an intermediate level of expression.

The ordering of the rows and columns was determined using hierarchical clustering and the associated dendrogram for the samples shown at the top of the figure. In this example, six relapsing patients were clustered together to the left and the non-relapsed patients clustered to the right. The heat map gives an overall view that the 348 genes have low expression in the relapse patients as indicated by the green color in the left-hand columns under the relapse patients. In

Figure 4. An example of a heat map. Reproduced from [23].



contrast, the non-relapse patients have higher expression levels for these genes as indicated by the red and black colors in their columns.

## Other methods

Our group has developed and used several other methods based on clustering. We will provide a brief description of three of these methods and provide references for detailed descriptions. We have found these methods useful with data from various studies including psoriasis [24], oncology [25] and pharmacogenetics [26].

### *K-means clustering and self-organizing maps*

Hierarchical clustering assumes a hierarchical structure in the data wherein all the genes start separately in their own cluster at the bottom of the dendrogram and iteratively merge into larger clusters as one goes up the tree. K-means and self-organizing maps (SOMs) cluster genes without assuming a hierarchical structure. Instead K-means starts with k genes sampled randomly from the data. Each of these genes is used as the starting center of one of the k clusters. Distances are calculated from each gene in the data to each of these *k* centers. Genes are then assigned to the closest center. Each center is then replaced by the average of the genes assigned to it. The algorithm repeats by recalculating the distance from each gene in the data to these new centers and reassigning genes to the closest center. This repeats until no genes are reassigned. For example, consider two gene clusters involved in the regulation of non-overlapping metabolic pathways. It may not be reasonable ever to merge these two groups of genes, as would be required by the hierarchical structure of a dendrogram. SOMs are similar to k-means but with slightly different iteration and update steps. The details of this technical distinction are not pertinent to this overview article [11,12,24,27,28].

### *Mantel statistic*

Mantel statistics provide a method to assess the correlation between two distance measures on the same data [29,30]. We applied this method to microarray data measured on brain tumors to statistically correlate the expression patterns with clinical covariates [22,25]. Since we wanted the clustering done on the samples, we used the transposed data, $X^T$. Two distance matrices, one based on the microarray expression values and the other using the clinical information, represent the pair-wise distances between the same samples in terms of two different factors. If samples that are far apart in the distance based on the microarray data are also far apart in the distance based on the clinical data and samples that are close in the microarray data are close in their clinical data, the pair-wise distances are positively correlated. This can indicate that clinical differences are related to gene expression differences. The Mantel statistic provides a formal statistical framework for quantifying these relationships and permutation tests can provide accurate p-values for testing significance.

### Consensus methods

This is a mathematical framework to combine the results of multiple cluster analyses into a final cluster result [22,31,32]. Conceptually, if two genes are very similar, they will be clustered together by most hierarchical clustering algorithms, distance measures and reasonable stopping rules. A consensus method will put those two genes into the same cluster in the final analysis. Similarly, if two genes seldom appear together, the consensus method will not put them in the same cluster. We are currently investigating how consensus methods might automate the choice of where to cut a dendrogram using bootstrapping to generate multiple cluster results and have applied it with encouraging results.

## Conclusion

We have focused on presenting an overview of hierarchical clustering of microarray data as a tutorial, emphasizing the relationship between a dendrogram and spatial representations of genes. We believe this relationship provides an intuitive understanding of how to analyze microarray data and can make it easier to interpret the results of a cluster analysis in a biological framework. The fact that the 'heat maps' found in the majority of microarray publications are based on hierarchical clustering indicates that an understanding of this general method is valuable to those who are just beginning to read the microarray literature and even to those who are using supervised methods.

We have avoided a discussion of implementation since most major statistical packages provide methods for cluster analysis and visualization and the choice of the package will depend on the level of computational and statistical expertise available in the particular lab. In our case, as professional statisticians, mathematicians and computer scientists, we use two advanced statistics packages: SAS (Cary, North Carolina) and S-Plus (Seattle, Washington). These packages contain many standard-clustering approaches used with microarray data and can be programmed to perform novel methods such as Mantel statistics or consensus methods. However, these packages require a high level of programming skill and most research groups will want to look for a statistical package that is easier to use.

We presented a brief description of hierarchical and some non-hierarchical clustering methods based on distance measures that our lab has used with success. There are many non-distance-based methods available, including principal component analysis, gene shaving, Bayesian methods and mixed-models approaches. We cannot present all of them in this review article and have not yet had the need to use them in our own work. Our view is that microarray data should be analyzed using distance-based methods instead of parametric model approaches because the assumptions for parametric models are currently hard to justify. We realize, of course, that as researchers gain more experience analyzing microarray data using parametric models and develop a solid probabilistic foundation for these approaches, some of these non-clustering methods may later become the *de facto* analytical framework of choice.

We emphasize the complexity and technical difficulty of performing cluster analysis. We do not see these methods as trivial to implement and would encourage researchers to begin building long-term collaborations with statisticians. However, cluster analysis and microarray data present novel problems with which many statisticians will have had no experience. Therefore, the collaboration will require a significant investment to introduce the statistician to these fields. One attraction to this field for a statistician is the opportunity for novel statistical methods research. This should be emphasized and supported as these collaborations develop.

Finally, we mention the Classification Society of North America [101] as an excellent cluster analysis resource. This organization supports the development of clustering and classification methods and the application of these methods to many academic fields. The society also publishes the prestigious *Journal of Classification* [102], which publishes fundamental papers on cluster analysis. In addition, the society maintains the *class-l* list server, an excellent forum for raising questions about cluster analysis. It is accessible through their web page [102].

## Outlook

Current methods of analyzing microarray data based on hierarchical clustering use 'off-the-shelf' algorithms developed over the last 50 years. Little work to date has been done to modify these methods for microarray data taking into account biological knowledge such as expected clusterings based on genes involved in metabolic pathways or genes sharing regulatory sites. Incorporating this type of knowledge will require a significant investment of time and support for statistical methodologists, but the added value of this research investment to pharmacogenomic studies should be huge.

### Highlights

- Supervised learning methods can predict membership in predetermined groups and identify genes important for classification. They require training data with known group assignment for each data point.
- Cluster analyses attempt to detect natural groups in data and identify genes important for classification. No *a priori* group assignments are required.
- Cluster analysis consists of a collection of distance-based unsupervised learning methods including hierarchical clustering, k-means clustering, self-organizing maps, principal components analysis, and Mantel statistics.
- Gene expression microarray data is typically filtered and normalized before using cluster analysis.
- Cluster analysis results should be used for data reduction and hypothesis generation.
- The heat map, a useful data visualization and summary tool, is a product of hierarchical clustering.
- Drawbacks of cluster analysis include lack of statistical tests for determining the number of clusters or the strength of cluster membership.

## Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Fisher R: The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7, 179-188 (1936).
2. Brown MP, Grundy WN, Lin D *et al.*: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97(1), 262-267 (2000).
3. Cristianini N, Shawe-Taylor J: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge (2000).
4. Bishop C: *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford (1996).
5. Khan J, Wei J, Ringner M *et al.*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673-679 (2001).
6. Ringner M, Peterson C, Khan J *et al.*: Analyzing array data using supervised methods. *Pharmacogenomics* 3(3), 403-415 (2002).
7. Everitt B, Rabe-Hesketh S: *The Analysis of Proximity Data.* John Wiley, New York City (1997).
8. Eisen M, Spellman P, Brown PO *et al.*: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95, 14863-14868 (1998).

• **This paper presents the first application of hierarchical clustering to microarray data and is a landmark publication.**

9. Hartigan J, Wong M: A k-means clustering algorithm. *Applied Statistics* 28, 100-108 (1979).
10. Kohonen T: The self-organizing map. *Proc. IEEE* 78, 1464-1479 (1990).
11. Timm N: *Applied Multivariate Analysis.* Springer, New York City (2002).
12. Hastie T, Tibshirani R *et al.*: *The Elements of Statistical Learning.* Springer, New York City (2001).
13. Legendre P, Legendre L: *Numerical Ecology.* Elsevier, New York City (1998).
14. Li C, Hung Wong W: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2(8), RESEARCH0032 (2001).

• **This paper takes a more traditional statistical modeling approach to improve the estimate of the genes expression and identify genes differentially expressed across sample groups. The model-based approach reduces the variability of low expression estimates, and provides a natural method of calculating expression values. The standard errors attached to expression values can be used to assess the reliability of downstream analysis.**

15. Wolfinger RD, Gibson G, Wolfinger ED *et al.*: Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8(6), 625-637 (2001).
16. Steele, Torrie: *Principles and Procedures of Statistics: a Biometrical Approach.* McGraw-Hill, New York City (1980).
17. Yang Y, Dudoit S *et al.*: Normalization for cDNA microarray data. Dept of Statistics Technical Report, University of California Berkeley (2001).
18. Durbin B, Hardin J *et al.*: A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18, S105-S110 (2002).
19. Hastie T, Tibshirani R, Eisen MB *et al.*: 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 1(2), 1-21 (2000).

• **This paper presents a novel approach for clustering microarray data which combines unsupervised and supervised learning methods. The method presented here also allows genes to belong to more than one cluster.**

20. Bittner M, Meltzer P, Chen Y *et al.*: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406(6795), 536-540 (2000).
21. Shannon W, Banks D: Combining classification trees using maximum likelihood estimation. *Stat. Med.* 18(6), 727-740 (1999).
22. Shannon WD, Watson MA, Perry A, Rich K: Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genet. Epidemiol.* 23(1), 87-96 (2002).

• **This paper is one of the first to focus on the problem of statistically correlating expression profiles with clinical covariates. The method presented here uses distance-based calculations like in hierarchical clustering and thus avoids the problem of distribution assumptions.**

23. Slonim DK: Transcriptional profiling in cancer: the path to clinical

pharmacogenomics. *Pharmacogenomics* 2(2), 123-136 (2001).
24. Bowcock AM, Shannon W, Du F *et al.*: Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Hum. Mol. Genet.* 10(17), 1793-1805 (2001).
25. Watson MA, Perry A, Budhjara V *et al.*: Gene expression profiling with oligonucleotide microarrays distinguishes World Health Organization grade of oligodendrogliomas. *Cancer Res.* 61(5), 1825-1829 (2001).
26. Zhang W, Shannon W *et al.*: Protein expression profiling to define CPT-11 therapy strategies in common cancers. (2002) (In Preparation).
27. Tamayo P, Slonim D, Mesirow J *et al.*: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96(6), 2907-2912 (1999).
28. Tavazoie S, Hughes JD, Campbell MJ *et al.*: Systematic determination of genetic network architecture. *Nat. Genet.* 22(3), 281-285 (1999).
29. Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27(2), 209-220 (1967).
30. Smouse P, Long J *et al.*: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* 35(4), 627-632 (1986).

## Websites

101. www.cs-na.org
Classification Society of North America.
102. http://link.springer-ny.com/link/service/journals/00357/
Journal of Classification.

BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 2 2005, pages 171–178
doi:10.1093/bioinformatics/bth469

# Outcome signature genes in breast cancer: is there a unique set?

Liat Ein-Dor[1,†], Itai Kela[1,3,†], Gad Getz[1,†], David Givol[2] and Eytan Domany[1,*]

[1]Department of Physics of Complex Systems, [2]Department of Molecular Cell Biology and [3]Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel

Received on June 4, 2004; revised on August 2, 2004; accepted on August 3, 2004
Advance Access publication August 12, 2004

## ABSTRACT

**Motivation:** Predicting the metastatic potential of primary malignant tissues has direct bearing on the choice of therapy. Several microarray studies yielded gene sets whose expression profiles successfully predicted survival. Nevertheless, the overlap between these gene sets is almost zero. Such small overlaps were observed also in other complex diseases, and the variables that could account for the differences had evoked a wide interest. One of the main open questions in this context is whether the disparity can be attributed only to trivial reasons such as different technologies, different patients and different types of analyses.

**Results:** To answer this question, we concentrated on a single breast cancer dataset, and analyzed it by a single method, the one which was used by van't Veer *et al.* to produce a set of outcome-predictive genes. We showed that, in fact, the resulting set of genes is not unique; it is strongly influenced by the subset of patients used for gene selection. Many equally predictive lists could have been produced from the same analysis. Three main properties of the data explain this sensitivity: (1) many genes are correlated with survival; (2) the differences between these correlations are small; (3) the correlations fluctuate strongly when measured over different subsets of patients. A possible biological explanation for these properties is discussed.

**Contact:** eytan.domany@weizmann.ac.il

**Supplementary information:** http://www.weizmann.ac.il/physics/complex/compphys/downloads/liate/

## INTRODUCTION

Several attempts were made to predict survival of cancer patients in general (Bair and Tibshirani, 2004; Beer *et al.*, 2002; Khan *et al.*, 2001; Nguyen and Rocke, 2002; Rosenwald *et al.*, 2002), and of breast cancer patients in particular (Ramaswamy *et al.*, 2003; Sorlie *et al.*, 2001; van't Veer *et al.*, 2002) on the basis of gene expression profiling. Sorlie *et al.* (2001) used an unsupervised approach, hierarchical clustering, to assign breast carcinoma tissues to one of five different subtypes, each with a distinctive expression profile. Robustness of these survival-related subclasses was demonstrated (Sorlie *et al.*, 2003) by applying the same analysis procedure to two independent breast carcinoma datasets (van't Veer *et al.*, 2002; West *et al.*, 2001). van't Veer *et al.* (2002) applied a supervised approach to identify a gene expression signature, based on 70 genes, capable of predicting a short interval to the development of distant metastases. First, they randomly selected a set of 78 patients, a training set, which was used to measure the correlation between each gene's expression and disease outcome. The genes were ranked according to this correlation, and the 70 most-correlated genes were used to construct a classifier discriminating between patients with good- and poor prognosis. The remaining 19 patients served as the test set to validate their prognosis classifier. A follow-up study (van de Vijver *et al.*, 2002) proved the efficiency of this classifier as a survival predictor on a large set of 295 tumor specimens. In a third study, Ramaswamy *et al.* (2003) identified a set of 128 genes separating metastases from primary tumors. A refined set of 17 metastases-associated genes were tested on a large diverse set of primary solid tumors, and were found to successfully distinguish patients with good versus poor prognosis.

The predictive success of these studies was frustrated by the fact that the sets of survival-related genes identified by these three studies had only a few genes in common. Only 17 genes appeared in both the list of 456 genes of Sorlie *et al.* (2001) and the 231 genes of van't Veer *et al.* (2002); merely 2 genes were shared between the sets of Sorlie *et al.* (2001) and Ramaswamy *et al.* (2003) Such disparity is not limited to breast cancer but characterizes other human disease datasets (Lossos *et al.*, 2004) such as schizophrenia (Miklos and Maleszka, 2004).

---

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

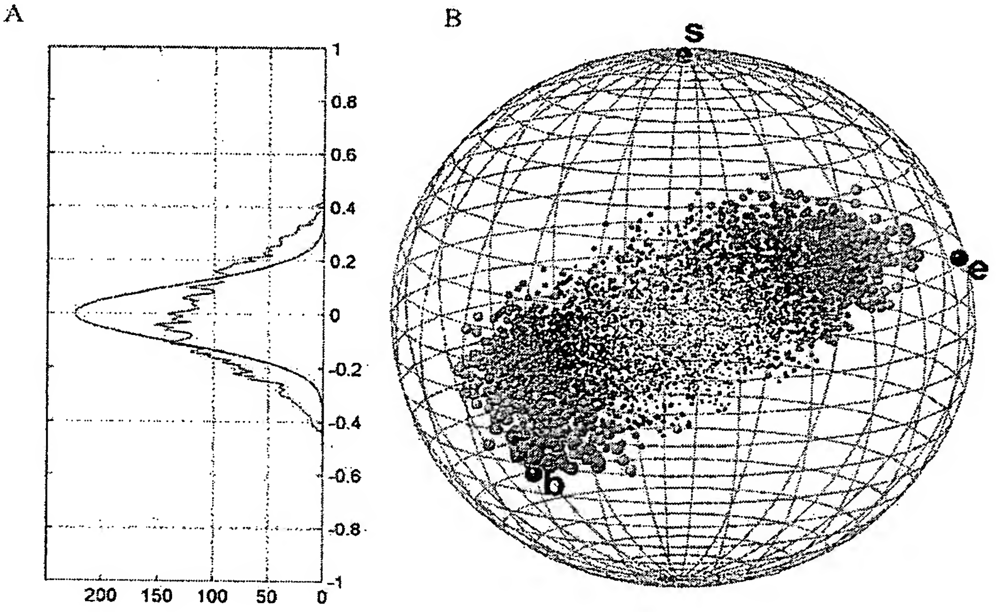


**Fig. 1.** (A) The histogram of the genes' correlation with the real survival vector (projection onto the vertical s axis—red curve), and with a random permutation of the survival vector (blue curve). (**B**) Globe of genes in the 'world' spanned by the normalized survival (**s**), BUB1 (**b**) and ESR1 (**e**). The survival is located at the north pole, while BUB1 (chosen from a large cluster of genes characterized by negative correlation with survival) and ESR1 (chosen from a large cluster of genes characterized by positive correlation with survival) are on the sphere's surface and their relative locations are determined by their angles with survival and with each other. All other (normalized) genes are represented by spots whose size and color illustrate how close the gene is to the surface (large red spots are close and small blue are far). The genes create an elongated structure at an angle $<\pi/2$ with s, implying that a large number of genes exhibit non-vanishing correlations with survival.

In this work, we explore this surprising phenomenon, and suggest new explanations for the lack of agreement between the sets of genes.

## MATERIALS AND METHODS

### Public dataset

The data van't Veer *et al.* (2002) contain gene expression profiles of primary breast tumors, from 96 sporadic young patients with grade T1/T2 tumors <5 cm in size, and N0 (no lymph node metastases). Of the 96 sporadic patients, 34 were treated by modified radical mastectomy and 62 underwent breast-conserving treatment, including axillary lymph node dissection followed by radiotherapy. Hybridization ratios were measured with respect to a reference made by pooling equal amounts of cRNA from all the sporadic carcinomas, on microarrays containing 25 000 human genes (Hughes *et al.*, 2001).

### Preprocessing of data

The full expression matrix of van't Veer *et al.* (2002) had 24 481 rows (genes) and 117 columns (samples). We applied filtering criteria, based on the entire set of 117 samples, yielding 5852 genes that exhibited a 2-fold change of expression with a $P$-value < 0.01 in five or more samples [van't Veer *et al.* (2002) applied the same filtering criteria on 98 samples, while discarding the test set of 19 samples, yielding 5000 genes]. We discarded from the set a single sample (sample 54) that contained >20% missing values [van't Veer *et al.* (2002) decided to include this patient in their analysis]. Like van't Veer *et al.* (2002) we also based our analysis on 96 'sporadic' patients free of BRCA1/2 germ line mutations.

### Correlation analysis

For each gene, we test the null hypothesis that its gene expression profile is uncorrelated with the survival vector (over all 96 samples). We randomly permuted the survival vector ($10^5$ times) and calculated the correlation of the expression of each gene with the randomized survival vector. The $P$-value is the fraction of times one gets an absolute correlation larger or equal to the absolute correlation of the unpermuted data. Correction for multiple comparisons was performed using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). Bounding the expected FDR by 10% yielded a list of 1234 genes for which the null hypothesis can be rejected. Histograms of the correlation (measured for 5852 genes) with the true survival and with a randomly permuted survival vector, are shown in Figure 1A.

### Dividing the data into ten different divisions of 77/19

To examine how different experiments of 77 samples influence the composition of the 70 most-correlated genes with survival, we used the bootstrapping method (Tibshirani, 1993). Bootstrapping is a computer simulation enabling the overcoming of finite size effects. It assumes that the sample is a good approximation of the population. By generating a large number of new samples from the original sample sets, we can estimate the statistical parameters of the population. To keep the good/poor prognosis ratio of the original training set (33/44) we divided the 96 samples into a poor prognosis set of 45 samples, and a good prognosis set of 51. We chose with repetitions a random set of 33 samples from the poor prognosis set, and 44 from the good prognosis. We repeated this procedure ten times and found the top 70 genes for each 'training set' composition.

### Measuring the STD of a gene based on a sample size of 77

We assumed that the degree of the polynomial fit for the average STD curve (Fig. 5) is the degree of the polynomial fit to the STD curve of each individual gene. Using this assumption, we found the polynomial fit to the STD curve of each gene in the data, and used it to estimate their STD values in a sample size of 77.

## RESULTS

### Many genes are related to survival

As was mentioned before, several microarray studies yielded gene sets whose expression profiles successfully predicted survival in breast cancer. However, the overlap between these gene sets was almost zero. This lack of agreement can be attributed to different chips, different methods of sample preparation, mRNA extraction and analysis of the data and, most importantly, to genuine differences between the patients (tumor grade, stage, etc.). To eliminate these sources of variation, we focused on data from a single experiment (van't Veer *et al.*, 2002). The data consist of 96 samples and 5852 genes (see Materials and Methods). Disease outcome is represented by a survival vector s, of 96 binary components, with 1 representing good prognosis (metastasis-free time interval >5 years), and 0 representing poor prognosis (<5 years). The projection of the 96-dimensional expression vector of each gene onto a three-dimensional space [spanned by the survival vector (s) and the expression vectors of ESR1 (e) and BUB1 (b)] is shown in Figure 1B.

We chose to use ESR1 and BUB1 as representative genes of two large clusters characterized by positive- and negative correlation with survival, respectively.

The 5852 genes comprise an oblate spheroid shaped cloud, tilted with respect to the equator. If survival is replaced by a random binary vector, the oblate spheroid cloud lies on the plane of the equator. Since the vertical component of each gene is the correlation of its expression with survival (Fig. 1A), this difference is a striking geometrical manifestation of the fact that the expression vectors of very many genes (1234—at an FDR of 10%, see Materials and Methods) are related to survival.

According to our model, if the experiment is repeated on a different group of patients (with the same clinical characteristics), the overall appearance of the new 'globe' will be quite similar, but the positions of individual genes will swarm around. This swarming will suffice to change drastically the relative ranking of the genes on the basis of their correlation with survival.

### Many sets of 70 genes can be used to predict survival

This dataset is characterized by three main properties: first, many genes are correlated with survival; second, the differences between these correlations are small; and third, the correlation-based rankings of the genes depend strongly on the training set (shown later). These properties may indicate that the top 70 genes are not superior to others in predicting disease outcome. To test this hypothesis, we selected the same 77 patients (out of 78; see Materials and Methods, and van't Veer *et al.*, 2002) and ranked all genes according to their correlation with survival. We used the 5852 genes to build a series of classifiers (following the method used by van't Veer *et al.*, 2002), based on consecutive groups of 70 genes. For each classifier, we measured the training and the test error, and found seven other sets of 70 genes, producing classifiers with the same prognostic capabilities as those based on the top 70. The genes of some of these seven classifiers appeared way down in the correlation-ranked list; the 70 genes of the first classifier are ranked between 71 and 140; classifier 2, 141–210; classifier 3, 211–280; classifier 4, 281–350; classifier 4, 351–420; classifier 5, 421–490; classifier 6, 561–630; classifier 7, 701–770. The location of these seven sets on the globe and their predicting performance is shown in Figures 9 and 11, respectively (see Supplementary information), and the corresponding Kaplan–Meier plots are shown below (Fig. 2).

To ensure that the aforementioned phenomenon is not unique to the specific training and test sets selected by van't Veer *et al.* (2002), we repeated the procedure described above for 1000 different compositions of training sets (of 77 samples) and test sets (19 samples). Each training set was used to rank the genes, and for each case the sequence of classifiers described above was constructed, and the training and test errors were measured for each classifier. Note, that when repeating this procedure for a randomized survival vector, the training error curve fluctuates around 37.5 mistakes (50% rate of errors) while the test error fluctuates around 9.5 mistakes, independent on the genes' rank. The results shown in Figure 3 imply that indeed, for each of the training sets, classifiers based on very-low-ranked genes are capable of predicting survival with quality similar to the high-ranking ones.

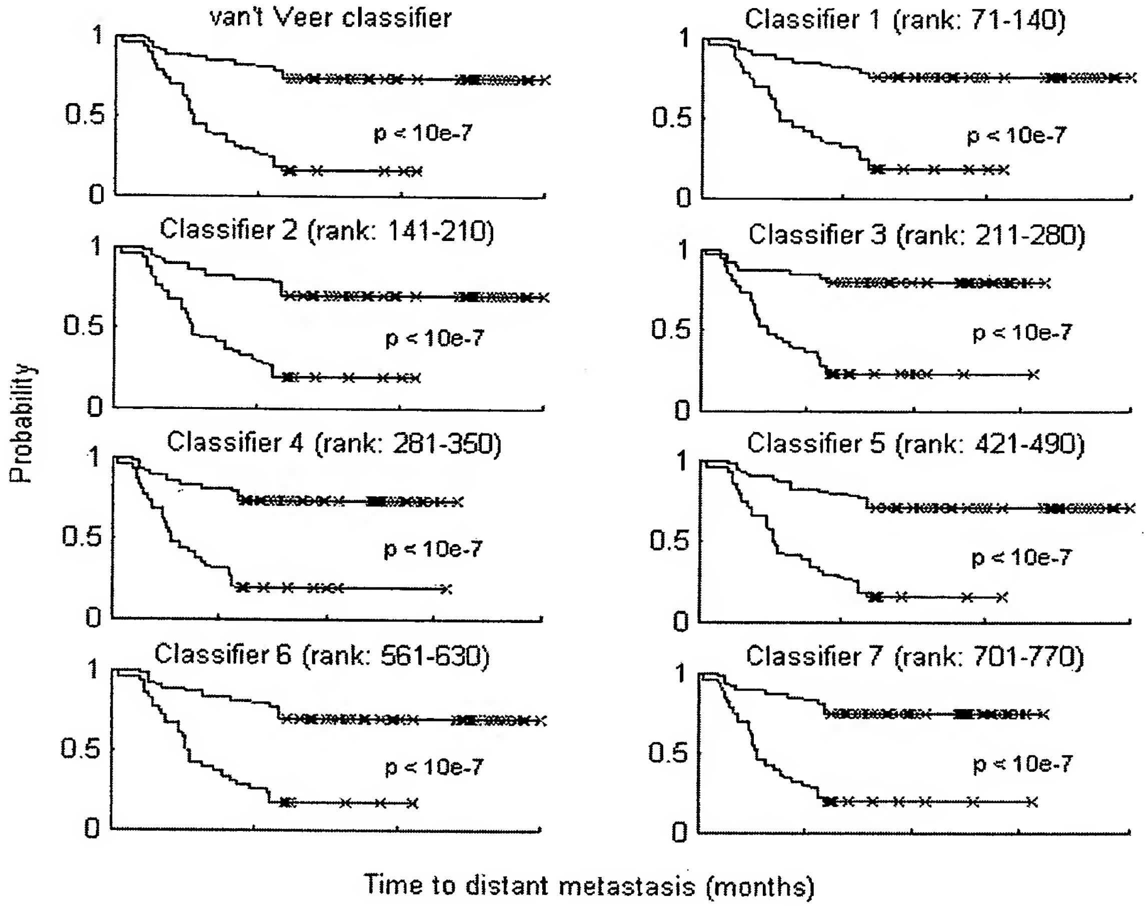


**Fig. 2.** Kaplan–Meier analysis of van't Veer *et al.*'s classifier and of the seven alternative classifiers as obtained from classifying all 96 samples. Upper curves describe the probability of remaining free of metastasis in the group of samples classified as having a good prognosis signature, while the lower curves describe the poor prognosis group.

To give a quantitative meaning to this claim, we generated the histogram presented in the inset of Figure 3, which shows that >70% of the 1000 training sets produced at least one classifier with the same (or better) performance as the one based on its own top 70 genes. The average number of such classifiers is four. The surprising summary of these observations is that (1) the list of the 'top 70 genes' of highest correlation with survival depends strongly on the training set of (77) patients on which the correlation was measured and (2) even with a fixed training set, one could have easily singled out a different group of 70 much lower ranked genes with as good a prognostic performance as that of the top-ranked genes.

Our results imply that although the top 70 genes may provide good prediction, other groups of 70 genes may do the same. Hence, these 70 genes cannot be considered as the main candidates for targeting anti-cancer treatment. Such candidates should be selected from the much longer list of genes related to survival, as demonstrated by the following list of cancer-related genes, present in the seven classifiers mentioned above. We list several of these genes, and indicate next to each one its correlation rank (in parentheses) measured on the training set selected by van't Veer *et al.* (2002).

*Negative correlation with survival* IL-6 (rank = 502) is anti-apoptotic, and therefore supports tumor survival (Lotem *et al.*, 2003); CDC25B (402) (Nilsson and Hoffmann, 2000), CKS2 (297) (Urbanowicz-Kachnowicz *et al.*, 1999), CDC2 (229) (Winters *et al.*, 2001) and CDC20 (341) (Singhal *et al.*, 2003) are known to function in cell cycle regulation or DNA replication; oncogenes NRAS (260) (Boon *et al.*, 2003) and EZH2 (92) (Varambally *et al.*, 2002) enhance cancer aggressiveness.

*Positive correlation with survival* It may be caused by some indirect relation to tumor growth, affecting survival through indirect mechanisms like immunity, apoptosis or inhibition of oncogenes. Examples: BIN1/AMPH2 (477) by binding to MYC functions as a tumor suppressor (Sakamuro *et al.*, 1996); BIK (342) is pro-apoptotic (Li *et al.*, 2003) via binding to BCL2 (1106) (Li *et al.*, 2003). The positive

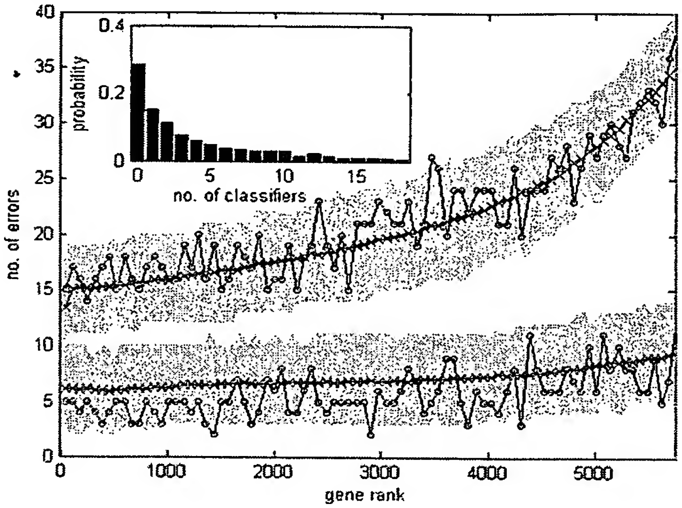


**Fig. 3.** The average performance of a series of classifiers generated by consecutive sets of 70 genes. The fluctuating curves present the number of errors produced by the classifiers resulting from one particular selection of training and test sets (upper, training errors out of 77 samples; lower, test errors out of 19). The $x$-axis represents the rank of the genes in the classifiers. The average over 1000 partitions is plotted as black x's; the two gray areas are the 95% confidence intervals of the training and test errors. Inset: histogram of the number of classifiers whose training and test errors are at least as low as those of the first classifier (based on the 70 genes with highest correlation to survival). Of the 1000 partitions, for ~28% no such classifier was found, whereas for ~6% five were found. Note that >70% of the training sets produce at least one classifier with the same performance as the top 70 genes; the expected number of such classifiers is around 4.

correlation of FLT3 (220) is due to its strong effect on dendritic cells and T-cells to enhance anti-tumor immunity (Ciavarra *et al.*, 2003). BRAK (237) is highly expressed in all normal tissues but low in malignant cells (Hromas *et al.*, 1999); IGFBP4 (225) induces apoptosis (Byron and Yee, 2003; Zhou *et al.*, 2003). Expression of GATA3 (255) is highly correlated with ER status (Bertucci *et al.*, 2000). Similarly, MYB (285) is also positively correlated with breast cancer outcome since it is a target of ER (Bertucci *et al.*, 2000; Guerin *et al.*, 1990) which is positively correlated with outcome. None of the genes listed above is ranked among the top 70.

Note that as opposed to claims made in (Gruvberger *et al.*, 2003), the success of the classifier is not due to the correlation of outcome to ER status. Creating a dataset which lacks this correlation, our seven classifiers, as well as van't Veer *et al.*'s (2002), kept their prognostic capabilities (see Supplementary information).

## A gene's rank may fluctuate

Say we measure the correlation $r$ of a gene's expression with survival on the basis of a sample of $N$ patients drawn at random from a larger group with similar clinical characteristics. If a different set of $N$ is drawn, the correlation will be different. If these statistical fluctuations of $r$ are sizeable, they may change the ranking of a gene from high in one sample to a much lower rank in another; the smaller the $N$, the larger the fluctuations of $r$. In order to estimate the effect of these fluctuations on the composition of gene lists such as those of van't Veer *et al.* (2002), we repeatedly selected different subgroups of 77 samples out of the 96 (in each group we maintained the overall good/poor prognosis ratio) and for each subgroup identified the 70 genes that have the highest correlation with survival. The significant variation of the membership of the top 70 genes is clearly shown in Figure 10 of the Supplementary information. Note that every pair of these training sets has at least 58 samples in common, which significantly reduces the fluctuations of $r$ and variation of the genes' ranks. In spite of this, the average overlap between two such gene groups is only 33.7/70. To better estimate the 'true' fluctuations of $r$ for independent subgroups of 77 we used bootstrapping (Tibshirani, 1993), drawing subgroups from the 96 samples with repeats (see Materials and Methods). This reduces the expected overlap of two top 70 gene lists to 12.2/70. Figure 4 shows how large the variation of gene rank is, measured for 10 subgroups. Genes whose correlation with survival ranked high over one subgroup are likely to become low ranked in another. Hence, different sets of 77 patients, drawn from a clinically similar pool, will yield different lists of 'top 70 genes' with respect to correlation with survival.

## Measuring the correlation fluctuations

In order to study how the fluctuations of the correlation with survival vary with the sample size $K$, we created $n_K$ non-overlapping subgroups of size $K$ from the 96 available samples. We calculated the correlation of each gene $g$ with survival, measured over each subgroup, and from these $n_K$ values we estimated the standard deviation (STD) of the correlation. We repeated this procedure five times (each time creating a different set of subgroups), to obtain $\sigma_g(K)$, the average STD, for each of the 5852 genes, for $K$ ranging from 2 to 48 (the maximal $K$ allowing for non-overlapping subgroups). Finally, we extrapolated the correlation noise (estimated by $\langle\sigma\rangle$, the STD averaged over the genes), from $K = 0$–96 (Fig. 5). As shown in Figure 5 the correlation noise decreases as the samples size increases. For sample size of 77 (the size of the training set), the expected average noise is ~0.1, whereas the significant genes found by van't Veer *et al.* (2002) and by our study show correlation between 0.3 and 0.5. In light of this small signal to noise ratio, the phenomenon shown in Figure 4 is not surprising.

Focusing on sample size $K = 77$ (Fig. 6), one can see that even relatively-low-ranked genes (around 1000), may have a non-negligible probability to be included among the 70 top ranked genes. Conversely, genes ranked among the top 70 can easily fluctuate to much lower ranks. The relatively low

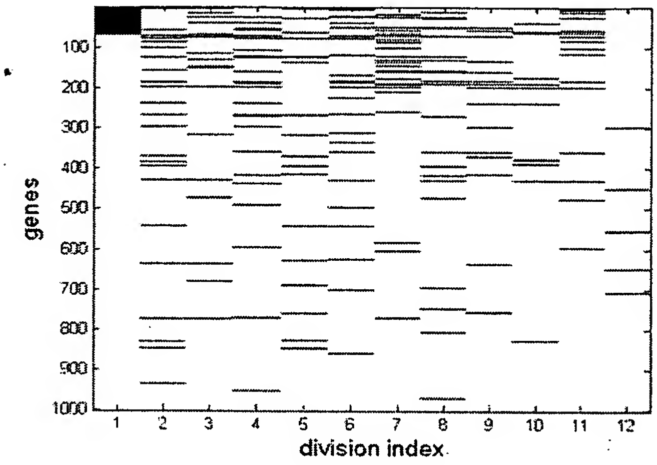


**Fig. 4.** Ten sets of top 70 genes, identified in 10 randomly chosen training sets of $N = 77$ patients (using bootstrapping—see Materials and Methods). Each row represents a gene and each column a training set. The genes were ordered according to their correlation rank in the first training set (leftmost column). For each training set, the 70 top-ranked genes are colored black. The genes that were top ranked in one training set can have a much lower rank when another training set is used. The two rightmost columns (columns 11 and 12) mark those of the 70 genes published by van't Veer *et al.* (2002) and the 128 genes appearing in (Ramaswamy *et al.*, 2003) that are among the top 1000 of our first training set.

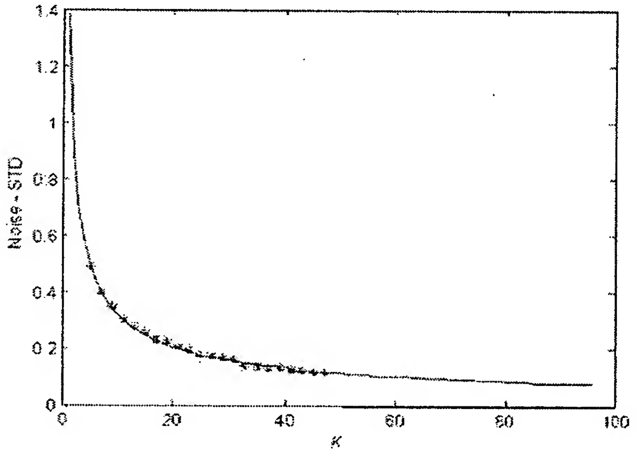


**Fig. 5.** Standard deviation (STD) of a gene's correlation with disease outcome, averaged over 5852 genes ($y$-axis) as a function of sample size K ($x$-axis). The curve is the polynomial fit to the results obtained for $K$ between 2 to 48. This curve was used to extrapolate the STD to larger values of $K$. (The values extrapolated to $K = 77$ were used to calculate the error-bars presented in Fig. 6.)

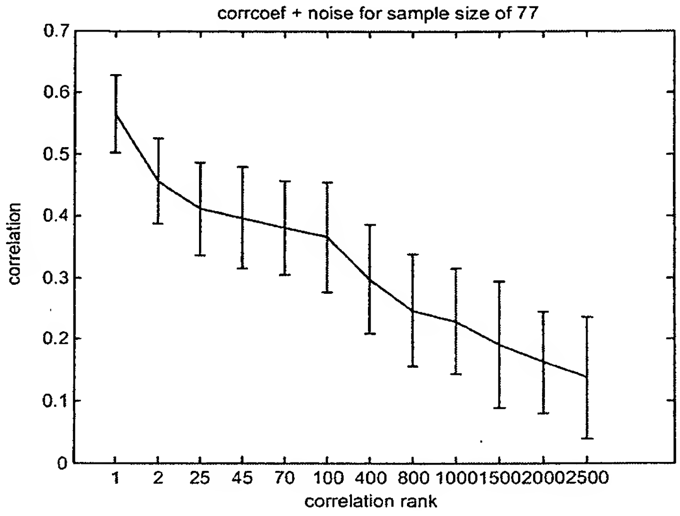


**Fig. 6.** Correlation of genes with survival versus their ranks. The correlation of each gene ($y$-axis) was measured based on the 96 samples, and the genes were ordered according to their correlation magnitude ($x$-axis). The error bars represent the noise (STD) of a gene based on sample size 77 (see Materials and Methods).

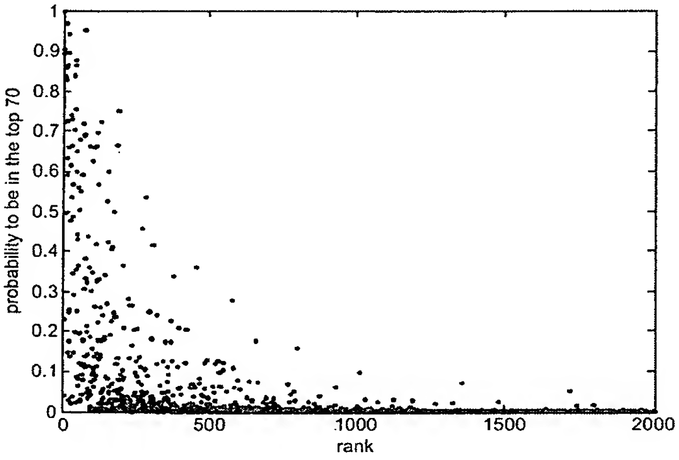


**Fig. 7.** The probability of genes to be included in a list of top 70. The genes were ranked on the basis of their correlation with outcome, as measured over the 77 samples of one particular (randomly chosen) training set.

signal-to-noise ratio explains the phenomenon demonstrated in Figure 4. In order to estimate the actual probability of each gene to be included in a list of top 70, we generated, at random, 10 000 training sets, each of 77 samples. For each such training set we identified the top 70 genes. The fraction of times (among 10 000) that each gene appeared in the top 70 is shown in Figure 7.

Taking into account the correlation noise, we defined an alternative gene score (instead of correlation coefficient), by calculating its probability to have a correlation above a given threshold for a given sample size (see Supplementary information). Figure 8 presents the probability of genes to

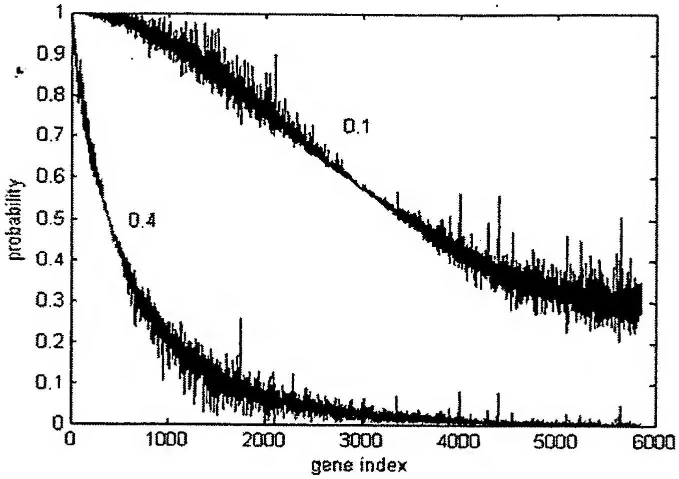


**Fig. 8.** The probability ($y$-axis) that genes have a correlation higher than a given threshold, calculated on the basis of noise derived for a training set of 77 samples. The $x$-axis represents the gene ranks according to their correlation with all 96 samples. The left curve corresponds to threshold of 0.4 and the right curve to threshold 0.1.

have a correlation higher than a given threshold ($y$-axis) calculated on the basis of the noise derived for a samples size of 77. The $x$-axis represents the genes' ranks according to their correlation coefficient with all 96 samples.

## DISCUSSION

In this work, we investigated a single breast cancer dataset (van't Veer *et al.*, 2002) in an attempt to explain the inconsistency between lists of survival-related genes derived from different experiments. While no single gene has a very high correlation with outcome, for many the correlation has intermediate values (Fig. 1). The differences between these correlation values are small, and the relative ranking of genes on the basis of correlation with survival changes drastically when a different training set is used. These large fluctuations in gene rank indicate that the identities of the top 70 ranked genes are not robust, and hence will not be reproduced in a different experiment. In spite of this sensitivity, the predictive power of several sets of genes is quite good. The main lesson is that whenever any arbitrary decision (e.g. choice of training and test set) is taken throughout analysis of the data, one has to generate a large ensemble of the different ways in which this arbitrary decision could be taken, and perform a statistical analysis of the results obtained over this ensemble. A high sensitivity of the results to the arbitrary decisions may indicate that the conclusions, e.g. the list of survival-related genes, are not unequivocal. In light of the inconsistency between lists of survival-related genes generated from the same dataset, the disagreement between lists obtained from different datasets is not surprising. A possible biological explanation for this may be the individual variations and heterogeneities associated with markers for outcome, even within a clinically homogenous group of patients.

Perhaps one has to divide the patients into smaller subgroups (Sorlie *et al.*, 2003) on the basis of some yet unknown attribute and for each subgroup of tumors look for it's much sought 'primary, master genes' that control the metastatic potential. The correlations with survival of such a master gene may be very high in its own subgroup and low in others. The large fluctuations in the correlation of such a gene's expression with survival, measured over different training sets, are due to the fluctuating fraction of how many members of the gene's subgroup are in the training set. It is important to note that such a master gene will not necessarily be top-ranked with respect to correlation measured in a very large sampling of patients, composed of a mixture of subgroups.

Since one may need much larger numbers of patients to identify such survival-wise-homogenous subgroups and their associated, potential master genes, one should separate two issues: the quest for survival-related master genes and the construction of prognostic tools on the basis of a short gene list. One can produce fairly reliable prognostic tools; many genes are related to survival, and using a large enough subset of them will compensate for the fluctuations in the predictive power of individual genes for individual patients. Membership in a prognostic list, however, is not necessarily indicative of the gene's importance in cancer pathology. Rather, in order to study the potential targets for treatment, one must scan the entire, wide list of survival-related genes. By focusing only on those genes that were singled out from one dataset as its preferred prognostic tool, one may miss important key players, in breast and also in other types of cancer.

## ACKNOWLEDGEMENTS

This work was supported by the Ridgefield Foundation, a Weizmann-Tel Aviv Sourasky Medical Center research grant, and by the NIH grant #5 P01 CA 65930-06.

## REFERENCES

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, E108.

Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.

Bertucci,F., Houlgatte,R., Benziane,A., Granjeaud,S., Adelaide,J., Tagett,R., Loriod,B., Jacquemier,J., Viens,P., Jordan,B., Birnbaum,D. and Nguyen,C. (2000) Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Hum. Mol. Genet.*, **9**, 2981–2991.

Boon,K., Edwards,J.B., Siu,I.M., Olschner,D., Eberhart,C.G., Marra,M.A., Strausberg,R.L. and Riggins,G.J. (2003) Comparison of medulloblastoma and normal neural transcriptomes identifies a restricted set of activated genes. *Oncogene*, **22**, 7687–7694.

Byron,S.A. and Yee,D. (2003) Potential therapeutic strategies to interrupt insulin-like growth factor signaling in breast cancer. *Semin. Oncol.*, **30**, 125–132.

Ciavarra,R.P., Brown,R.R., Holterman,D.A., Garrett,M., Glass,W.F.,II, Wright,G.L.,Jr, Schellhammer,P.F. and Somers,K.D. (2003) Impact of the tumor microenvironment on host infiltrating cells and the efficacy of flt3-ligand combination immunotherapy evaluated in a treatment model of mouse prostate cancer. *Cancer Immunol. Immunother.*, **52**, 535–545.

Gruvberger,S.K., Ringner,M., Eden,P., Borg,A., Ferno,M., Peterson,C. and Meltzer,P.S. (2003) Expression profiling to predict outcome in breast cancer: the influence of sample selection. *Breast Cancer Res.*, **5**, 23–26.

Guerin,M., Sheng,Z.M., Andrieu,N. and Riou,G. (1990) Strong association between c-myb and oestrogen-receptor expression in human breast cancer. *Oncogene*, **5**, 131–135.

Hromas,R., Broxmeyer,H.E., Kim,C., Nakshatri,H., Christopherson,K.,II, Azam,M. and Hou,Y.H. (1999) Cloning of BRAK, a novel divergent CXC chemokine preferentially expressed in normal versus malignant cells. *Biochem. Biophys. Res. Commun.*, **255**, 703–706.

Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.

Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Li,Y.M., Wen,Y., Zhou,B.P., Kuo,H.P., Ding,Q. and Hung,M.C. (2003) Enhancement of Bik antitumor effect by Bik mutants. *Cancer Res.*, **63**, 7630–7633.

Lossos,I.S., Czerwinski,D.K., Alizadeh,A.A., Wechser,M.A., Tibshirani,R., Botstein,D. and Levy,R. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med.*, **350**, 1828–1837.

Lotem,J., Gal,H., Kama,R., Amariglio,N., Rechavi,G., Domany,E., Sachs,L. and Givol,D. (2003) Inhibition of p53-induced apoptosis without affecting expression of p53-regulated genes. *Proc. Natl Acad. Sci. USA*, **100**, 6718–6723.

Miklos,G.L. and Maleszka,R. (2004) Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615–621.

Nguyen,D.V. and Rocke,D.M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Nilsson,I. and Hoffmann,I. (2000) Cell cycle regulation by the Cdc25 phosphatase family. *Prog. Cell Cycle Res.*, **4**, 107–114.

Ramaswamy,S., Ross,K.N., Lander,E.S. and Golub,T.R. (2003) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, **33**, 49–54.

Rosenwald,A., Wright,G., Chan,W.C., Connors,J.M., Campo,E., Fisher,R.I., Gascoyne,R.D., Muller-Hermelink,H.K., Smeland,E.B., Giltnane,J.M. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Sakamuro,D., Elliott,K.J., Wechsler-Reya,R. and Prendergast,G.C. (1996) BIN1 is a novel MYC-interacting protein with features of a tumour suppressor. *Nat. Genet.*, **14**, 69–77.

Singhal,S., Amin,K.M., Kruklitis,R., DeLong,P., Friscia,M.E., Litzky,L.A., Putt,M.E., Kaiser,L.R. and Albelda,S.M. (2003) Alterations in cell cycle genes in early stage lung adenocarcinoma identified by expression profiling. *Cancer Biol. Ther.*, **2**, 291–298.

Sorlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Sorlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R., Geisler,S. *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Tibshirani,B.E.a.R.J. (ed.) (1993) *An Introduction to the Bootstrap*. Chapman and Hall, NY.

Urbanowicz-Kachnowicz,I., Baghdassarian,N., Nakache,C., Gracia,D., Mekki,Y., Bryon,P.A. and Ffrench,M. (1999) ckshs expression is linked to cell proliferation in normal and malignant human lymphoid cells. *Int. J. Cancer*, **82**, 98–104.

van de Vijver,M.J., He,Y.D., van't Veer,L.J., Dai,H., Hart,A.A., Voskuil,D.W., Schreiber,G.J., Peterse,J.L., Roberts,C., Marton,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Varambally,S., Dhanasekaran,S.M., Zhou,M., Barrette,T.R., Kumar-Sinha,C., Sanda,M.G., Ghosh,D., Pienta,K.J., Sewalt,R.G., Otte,A.P., Rubin,M.A. and Chinnaiyan,A.M. (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624–629.

West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J.A.,Jr, Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Winters,Z.E., Hunt,N.C., Bradburn,M.J., Royds,J.A., Turley,H., Harris,A.L. and Norbury,C.J. (2001) Subcellular localisation of cyclin B, Cdc2 and p21(WAF1/CIP1) in breast cancer. association with prognosis. *Eur. J. Cancer*, **37**, 2405–2412.

Zhou,R., Diehl,D., Hoeflich,A., Lahm,H. and Wolf,E. (2003) IGF-binding protein-4: biochemical characteristics and functional consequences. *J. Endocrinol.*, **178**, 177–193.